Tampere University

GUANQUN CAO

Multi-view Data Analysis

Tampere University Dissertations 6

Tampere University Dissertations 6

GUANQUN CAO

Multi-view Data Analysis

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty Council of the Faculty of Computing and Electrical Engineering
of Tampere University of Technology,
for public discussion in the Auditorium TB109
of Tietotalo Building, Korkeakoulunkatu 1, Tampere,
on January 9th, 2019, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Supervisors	Moncef Gabbouj, Professor Tampere University Finland	Alexandros Iosifidis, Associate Professor Aarhus University Denmark
Pre-examiners	Guoying Zhao, Professor University of Oulu Finland	Abdulmotaleb El Saddik, Professor University of Ottawa Canada
Opponent	Gaurav Sharma, Professor University of Rochester USA	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-0967-1 (print)
ISBN 978-952-03-0968-8 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-0968-8

PunaMusta Oy
Tampere 2019

Abstract

Multi-view data analysis is a key technology for making effective decisions by leveraging information from multiple data sources. The process of data acquisition across various sensory modalities gives rise to the heterogeneous property of data. In my thesis, multi-view data representations are studied towards exploiting the enriched information encoded in different domains or feature types, and novel algorithms are formulated to enhance feature discriminability. Extracting informative data representation is a critical step in visual recognition and data mining tasks. Multi-view embeddings provide a new way of representation learning to bridge the semantic gap between the low-level observations and high-level human comprehensible knowledge benefitting from enriched information in multiple modalities.

Recent advances on multi-view learning have introduced a new paradigm in jointly modeling cross-modal data. Subspace learning method, which extracts compact features by exploiting a common latent space and fuses multi-view information, has emerged proiminent among different categories of multi-view learning techniques. This thesis provides novel solutions in learning compact and discriminative multi-view data representations by exploiting the data structures in low dimensional subspace. We also demonstrate the performance of the learned representation scheme on a number of challenging tasks in recognition, retrieval and ranking problems.

The major contribution of the thesis is a unified solution for subspace learning methods, which is extensible for multiple views, supervised learning, and non-linear transformations. Traditional statistical learning techniques including Canonical Correlation Analysis, Partial Least Square regression and Linear Discriminant Analysis are studied by constructing graphs of specific forms under the same framework. Methods using non-linear transforms based on kernels and (deep) neural networks are derived, which lead to superior performance compared to the linear ones. A novel multi-view discriminant embedding method is proposed by taking the view difference into consideration. Secondly, a multi-view nonparametric discriminant analysis method is introduced by exploiting the class boundary structure and discrepancy information of the available views. This allows for multiple projecion directions, by relaxing the Gaussian distribution assumption of related methods. Thirdly, we propose a composite ranking method by keeping a close correlation

with the individual rankings for optimal rank fusion. We propose a multi-objective solution to ranking problems by capturing inter-view and intra-view information using autoencoder-like networks. Finally, a novel end-to-end solution is introduced to enhance joint ranking with minimum view-specific ranking loss, so that we can achieve the maximum global view agreements within a single optimization process.

In summary, this thesis aims to address the challenges in representing multi-view data across different tasks. The proposed solutions have shown superior performance in numerous tasks, including object recognition, cross-modal image retrieval, face recognition and object ranking.

Preface

The research presented in this thesis has been carried out at the Laboratory of Signal Processing, Tampere University of Technology (TUT), Finland, during 2012-2018.

This thesis owes its existence to the help, support, and inspiration of many people. First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Moncef Gabbouj for his advice, support and patience over the years to make the thesis happen. Prof. Gabbouj has provided a wonderful environment which enables me to focus on the research work. His words of wisdom, critical thinking, and admirable personality conveyed through our discussions and meetings will benefit me in the long run.

I am indebted to Prof. Alexandros Iosifidis as my instructor, for his excellent guidance and continuous encouragement for my research work. I am very grateful for his precise instruction, relentless support, and careful review of my papers and thesis. I wish to express my gratitude to Dr. Ke Chen, whom I had the initial discussion about multi-view data analysis with. His inspiration and support have a deep impact for me to carry on the later research.

I want to thank Profs. Vijay Raghavan and Raju Gottumukkala for hosting me at University of Louisiana at Lafayette, in April 2017. I enjoyed our fruitful discussion and quality stay in Lafayette and very glad about the successful outcome from our collaboration.

I would like to thank the pre-examiners Prof. Guoying Zhao from University of Oulu, Finland and Prof. Abdulmotaleb El Saddik from The University of Ottawa for their valuable comments on the thesis.

I would like to thank Virve Larmila, Ulla Siltaloppi and Elina Orava, for their great help of routine but important administrative work.

I wish to extend my thanks to the members of the multimedia group, whom I spent most of my time while conducting the research. I want to thank the initial crew of office TC 413, including Dr. Stefan Uhlmann, Dr. Murat Birici and Dr. Jenni Raitoharju, for their discussions and suggestions on numerous topics in work and life. Special thank goes to Mr. Honglei Zhang, who generously grant me of his time and efforts for my work, career and life. The current members of TC 413 and the rest of the group are also

acknowledged.

I owe a special gratitude to my parents, for their patience and mental support through all the hardship during my doctoral study.

Finally, I wish to thank everyone who make a positive contribution to my thesis.

Contents

List of Figures

List of Tables

List of Symbols

$\mathbf{1}$	An all-ones matrix
c_i	Class label of the sample $\mathbf{x}_i$
C	Number of class
$\mathbf{e}$	A vector of ones
$\mathcal{G}$	A weighted graph
$\mathbf{I}$	Identity matrix
$\mathbf{K}$	Representation in the kernel space
$\mathbf{L}_B$	Between-class Laplacian matrix
$\mathbf{L}_W$	Within-class Laplacian matrix
$\mathbf{P}$	Inter-view covariance matrix
$\mathbf{Q}$	Intra-view covariance matrix
$\mathbb{R}$	A set of real numbers
$\mathbf{W}$	Data projection matrix
$\mathbf{W}_v$	Data projection matrix of the vth view
$\mathbf{x}$	A feature vector of D dimensions
$\mathbf{X}$	A data matrix of D dimensions by N samples
$\mathbf{X}_v$	Feature vectors of the vth view
d	Dimensionality in the latent space
λ	Eigenvectors
ϕ	Nonlinear mapping function to the Hilbert space
$\mathbf{\Sigma}$	Covariance matrix

List of Abbreviations

CCA	Canonical Correlation Analysis
GMA	Gerneralized Multi-view Analysis
KMvPLS	Kernel Multi-view Partial Least Square regressions
KapMvCCA	Approximate Kernel Multi-view Canonical Correlation Analysis
KMvCCA	Kernel Multi-view Canonical Correlation Analysis
KapMvPLS	Approximate Kernel Multi-view Partial Least Square regressions
LMvCCA	Linear Multi-view Canonical Correlation Analysis
LMvPLS	Linear Multi-view Partial Least Square regressions
MvCCAE	Multi-view Canonically Correlated Auto-Encoder
MvDA	Multi-view Discriminant Analysis
MULDA	Multi-view Uncorrelated Linear Discriminant Analysis
MvMDAE	Multi-view Modularly Discriminant Auto-Encoder
MvNDA	Multi-view Nonparametric Discriminant Analysis
RMSE	Root Mean Square Error

List of Publications

The thesis is composed of a summary part and 4 publications listed below as appendices. The publications are referred in the thesis as [P1], [P2], etc.

[P1] G. Cao, A. Iosifidis, K. Chen and M. Gabbouj, "Generalized Multi-View Embedding for Visual Recognition and Cross-Modal Retrieval," in IEEE Transactions on Cybernetics, vol. 48, no. 9, pp. 2542-2555, Sept. 2018. doi: 10.1109/TCYB.2017.2742705

[P2] G. Cao, A. Iosifidis and M. Gabbouj, "Multi-View Nonparametric Discriminant Analysis for Image Retrieval and Recognition," in IEEE Signal Processing Letters, vol. 24, no. 10, pp. 1537-1541, Oct. 2017. doi: 10.1109/LSP.2017.2748392

[P3] G. Cao, M. A. Waris, A. Iosifidis and M. Gabbouj, "Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, 2016, pp. 1-6. doi: 10.1109/IPTA.2016.7821032

[P4] G. Cao, A. Iosifidis, M. Gabbouj, V. Raghavan, R. Gottumukkala, Deep Multi-view Learning to Rank, submitted to IEEE Trans. on Knowledge and Data Engineering. arXiv:1801.10402

1 Introduction

1.1 Objective

The goal of the thesis is to analyze and learn data representations in different domains or feature types. In particular, the aim is to improve the performance of multi-view data analysis in numerous applications by exploiting the enriched information coming from various sources. Different groups of feature vectors are considered as views, and a view therefore includes feature vectors in a specific domain/modality, e.g. image, text or speech, or from a dedicated descriptor, e.g. color descriptor, texture descriptor, shape descriptor or audio descriptor. Cross-modal matching is also enabled on heterogeneous data, where direct matchings of data samples across feature spaces is infeasible. This is in contrast to applications like text-based image retrieval, which largely take advantage of ground truth textual labels surrounding the images. Although these applications are relatively successful under certain circumstances, the deficiency in developing the cross-modal learning capability limits them in fully understanding images.

Extracting informative data representations is a critical step in visual recognition and data mining tasks. It aims to bridge the semantic gap between the low-level feature representations and high-level human comprehensible knowledge. This thesis introduces several techniques in learning multi-view data representations using subspace learning techniques, and formulates novel algorithms to enhance feature discriminability. The learned feature representation provides a discriminative input for the future tasks. Moreover, end-to-end solutions are formulated to strenghten the learning capabability by optimizing the entire system thoroughly. To this end, the performance of many challenging problems in visual recognition and data mining is improved.

1.2 Motivation

Our intuition is that, visual objects are described from various view points and/or modalities. The process of identifying an object can not only benefit from visual descriptions, but interactions with image captions and enriched attributes [2]. With a huge volume of data generated from sensor technologies, visual recognition and data mining problems urge

people to dive into multi-view and cross-domain learning [3, 4]. Meanwhile, the process of data acquisition across diverse domains or representations by using different types of feature extraction methods give rise to the heterogeneous property of data. Therefore, in order to analyze the heterogeneous data, multi-view and cross-modal learning algorithms are formulated to significantly improve the performance of machine learning [4, 3]. The research of multi-view data analysis is of great importance, and is able to improve the performance of numerous applications, which include but are not limited to

- **Cross-modal Multimedia Retrieval.** This application uses queries from one domain (e.g. text) while seeking for similar contents from another domain (e.g. image). Examples can be found in [5, 3, 6].
- **Object Recognition.** With high-level semantic information embedded in the recognition model, the object recognition performance can be improved [7, 8].
- **Face Photo-Sketch Recognition.** The cross-modal matching between faces and sketches is made possible when their features are embedded in the shared subspace. Examples can be found in [9, 10].
- **Multi-view Learning to Rank.** The traditional ranking model evaluates the relevance between every pair of query and data by combining the features in an optimal way. In contrast, the relevance of the same pairs may differ from various ranking sources. A potential compositive ranking is provided as the solution to maximize the global agreement.
- **Visual question answering.** Two models are built to encode the visual and language views. Image and question embeddings are combined to obtain a single model, so that the visual question answering can be achieved [11].

1.3 Thesis Outline

The rest of the thesis is organized as follows. Chapter 2 reviews the literature of subspace learning, regularization and in particularly multi-view learning. The thesis contribution is presented in details in Chapter 3. A unified formulation to generalize the multi-view subspace learning methods is introduced. It is then extended to nonlinear mappings using kernels and neural networks. A new regularization scheme form linear multi-view subspace learning is described to prevent overfitting. A nonparametric multi-view learning technique is also introduced, which enables multiple projection directions, by relaxing the Gaussian distribution assumption of related methods. In the end, a composite ranking method from multiple sources is proposed to enhance the joint ranking and minimize the view-specific ranking loss. A schematic diagram illustrating the relation between methods in the thesis is shown in Figure 1.1.

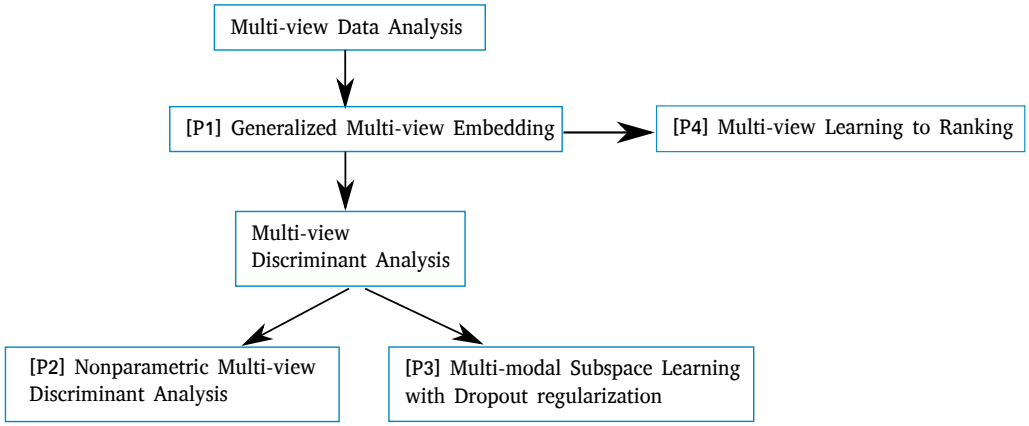


Figure 1.1: The schematic view of the thesis outline.

1.4 Contributions and Publications

The major contribution of this thesis is the proposal of a unified framework of multi-view data mapping, which is described in [P1]. A nonparametric extension is introduced in [P2], and further extension with Dropout-like regularization can be found in [P3]. Moreover, a multi-view learning to rank method, as an importance application of multi-view data analysis, is developed and described in [P4].

In [P1], an unified solution for subspace learning methods is proposed using the Rayleigh quotient, which is extensible for multiple views, supervised learning, and nonlinear embeddings. The proposed framework is generalized to numerous statistical learning methods including Canonical Correlation Analysis, Partial Least Square regression and Linear Discriminant Analysis with graphs in specific forms. It is also extented to nonlinear mappings using kernel and neural network-based methods. A simple yet effective Multi-view Modular Discriminant Analysis is proposed by introducing the view difference. The proposed multi-view embedding methods have shown superior performance in visual object recongition and cross-modal multimedia retrieval. The candidate is the first author of this publication and is responsible for developing most of the methods, performing all experiments and writing the manuscript.

In [P2], a novel multi-view nonparametric discriminant analysis method is proposed and achieves superior performance in cross-modal image retrieval and zero-shot recognition. The class boundary structure and view discrepancy is exploited to formulate an optimization criterion which is automatically adjusted to the multi-view class structures. The advantage of the new method is that it enables multiple projection directions, by relaxing the Gaussian distribution assumption of related methods. A better class discrimination is obtained using the new graph formulation leading to an improved performance. The candidate is the first author of this publication and is responsible for developing the whole methods, performing all experiments and writing the manuscript.

In [P3], inspired by the regularization for neural networks, a novel regularizer is introduced to artificially remove the effect of certain amount of feature bins using the probabilistic approach to prevent linear multi-view subspace learning from overfitting. A joint dropout-regularized multi-modal subspace learning algorithm is formulated which integrates within-class similarities and between-class separabilties to achieve good class separation. The objective function can be solved efficiently, and the method demontrates its effectiveness in face-sketch recognition and cross-modal retrieval problems. The candidate is the first author of this publication and is responsible for developing the whole methods, performing all experiments and writing the manuscript.

In [P4], a multi-view learning to rank is developed, which is one of the few methods in data mining. A composite ranking method is introduced to keep a close correlation with the individual rankings. Multi-objective solutions to ranking is devised by capturing the information of the feature mapping from both within each view as well as across views using autoencoder-like networks. Moreover, we introduce an end-to-end solution to enhance the joint ranking with minimum view-specific ranking loss, so that the maximum global view agreement is achieved in a single optimization process. Superior ranking results are achieved on university ranking, multi-view lingual text ranking and image data ranking problems. The candidate is the first author of this publication and is responsible for developing the whole methods, performing all experiments and writing the manuscript.

2 Related Work

In this chapter, we firstly review the subspace learning algorithms. Then, several important regularization methods are reviewed. Relevent methods in multi-view learning are elaborated. Finally, some applications of interest are provided.

2.1 Subspace Learning

Subspace learning is an important data analysis approach which is used to extract salient features from data. The main idea is to project the high-dimensional data into the low-dimensional space by fitting certain criteria, so that the relevant information to the subsequent processing is maintained [12, 13]. This type of approaches can be classified into three categories, based on the availability of class labels: unsupervised methods, supervised methods and semi-supervised methods. Unsupervised methods learn the underlying data patterns by using the similarities between samples. Traditional methods like principal component analysis (PCA) belong to this category. Supervised methods are effective in extracting discriminative features from the labeled data, and thus leading to good results in classification. Linear discriminant analysis (LDA) as the most representative supervised method, shows superior results in face recognition compared to PCA [14]. Semi-supervised methods make use of both labeled and unlabeled data, e.g. semi-supervised discriminant analysis (SDA) extends the objective function of LDA by using a graph-based regularization term. Many subspace learning methods can be described as specific cases of the graph embedding framework [15]. We describe these techniques in details in following sections.

2.1.1 Graph Embedding

Graph embedding has been considered as a general framework for dimensionality reduction [15, 16]. The basic idea is to find a mapping function $\mathbf{F} : \mathbf{X} \in \mathbb{R}^{D\times N} \rightarrow \mathbf{Y} \in \mathbb{R}^{d\times N}$ to map the data from the original high-dimensional space to a low-dimensional space, where $D > d$ is the dimensionality of the feature space and N is the number of samples. The function $\mathbf{F}$ can be linear or nonlinear, implicit or explicit, depending on the method used to define the data projection. The method assumes that we can develop

a weighted graph $\mathcal{G} = \{\mathbf{X}, \mathbf{V}\}$ with similarity matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$ over the training data $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$. A graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{V}$ can then be defined, where

$$\mathbf{D}[i,i] = \sum_j \mathbf{V}_{ij}, \text{ and } \mathbf{D}[i,j] = 0 \text{ for } \forall i \neq j. \tag{2.1}$$

We also define a penalty graph $\mathcal{G}^p = \{\mathbf{X}, \mathbf{V}^p\}$ formed by the same vertices $\mathbf{X}$ but using a different similarity weight matrix $\mathbf{V}^p$. The data is projected to a latent space, where the similarity charateristics of $\mathbf{V}^p$ are suppressed. Considering the sample-wise projection to $\mathbf{y} = [y_1, y_2, \cdots, y_N] \in \mathbb{R}^d$, the graph-preserving objective is

$$\mathbf{W}^* = \underset{\mathrm{Tr}(\mathbf{y}\,\mathbf{C}\mathbf{y}^\top)=q}{\arg\min} \sum_{i \neq j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \mathbf{V}_{ij} = \underset{\mathrm{Tr}(\mathbf{y}\,\mathbf{C}\mathbf{y}^\top)=q}{\arg\min}\, \mathbf{y}\,\mathbf{L}\mathbf{y}^\top, \tag{2.2}$$

where q is a constant and $\mathbf{C}$ is a matrix used to constrain the minimization of the objective function. $\mathbf{C}$ is also a diagonal matrix for scale optimization, and can also be the Laplacian matix of a penalty graph $\mathcal{G}^p$. In the latter case, $\mathbf{C} = \mathbf{L}^p = \mathbf{D}^p - \mathbf{V}^p$. Similar to (2.1), $\mathbf{D}^p$ is the diagnal matrix. The graph-preserving objective is the criterion of graph embedding for all vertices. While the graph vertices in direct graph embeddings only presents the training data, it can be extended to new test data in the original feature space.

The mapping $\mathcal{F} : \mathbf{X} \rightarrow \mathbf{Y}$ defined by 2.2 can take three forms:

Linear: We consider linear projections of the original data $\mathbf{x}_i \in \mathbb{R}^D$ to a low-dimensional feature space $\mathbb{R}^d$, $d < D$, which is expressed as $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$, where $\mathbf{W} \in \mathbb{R}^{D \times d}$ is the data projection matrix. The objective function in (2.2) can be expressed as

$$\mathbf{W}^* = \underset{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X}\mathbf{C}\mathbf{X}^\top \mathbf{W})=q}{\arg\min} \sum_{i \neq j} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 \mathbf{V}_{ij} = \underset{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X}\mathbf{C}\mathbf{X}^\top \mathbf{W})=q}{\arg\min} \mathbf{W}^\top \mathbf{X}\mathbf{L}\mathbf{X}^\top \mathbf{W}, \tag{2.3}$$

Kernel-based: The idea behind kernel methods is to map the data from the original feature space $\mathbb{R}^D$ to a higher dimensional Hilbert space $\mathcal{F}$. Let us define $\phi(\cdot)$ as the nonlinear function mapping $\mathbf{x}_i \in \mathbb{R}^D$ to $\mathcal{F}$, and $\mathbf{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]$ as the data matrix in $\mathcal{F}$. The kernel trick [17] is exploited in order to implicitly map the data to arbitrary space $\mathcal{F}$, and the kernel matrix $\mathbf{K} = \mathbf{\Phi}^\top \mathbf{\Phi}$ contains the inner products between the training samples in the Hilbert space, which can also be written as

$$[\mathbf{K}]_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)] = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j), \tag{2.4}$$

where $\kappa(\cdot,\cdot)$ is the so-called kernel function. The centered Gram matrix is $\bar{\mathbf{K}} = \mathbf{K} - \frac{1}{N}\mathbf{1}\,\mathbf{K} - \frac{1}{N}\mathbf{K}\,\mathbf{1}^\top + \frac{1}{N^2}\mathbf{1}\mathbf{K}\,\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{N \times N}$ is an all-ones matrix. In order to find the optimal projection, we can express $\mathbf{W}$ of each view as a linear combination of the training samples in the kernel space based on the Representer Theorem [18, 19]. The data projecttion matrix can be expressed by using a new weight matrix $\mathbf{A}$ as

$$\mathbf{W} = \mathbf{\Phi}\mathbf{A}. \tag{2.5}$$

The feature mapping in kernel methods can be derived as

$$\mathbf{Y} = \mathbf{W}^\top \mathbf{\Phi} = \mathbf{A}^\top \mathbf{\Phi}^\top \mathbf{\Phi} = \mathbf{A}^\top \mathbf{K}. \tag{2.6}$$

The objective function in (2.2) is written as

$$\mathbf{A}^* = \underset{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\mathbf{C}\mathbf{K}\mathbf{A})=q}{\arg\min} \sum_{i \neq j} \|\mathbf{A}^\top \mathbf{k}_i - \mathbf{A}^\top \mathbf{k}_j\|^2 \mathbf{V}_{ij} = \underset{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\mathbf{C}\mathbf{K}\mathbf{A})=q}{\arg\min} \mathbf{A}^\top \mathbf{K}\,\mathbf{L}\,\mathbf{K}\mathbf{A} \tag{2.7}$$

Neural networks-based: The data mappnig $\mathcal{F}$ can take the form of a neural network with M layers, where $\boldsymbol{\theta}_j$ contains the weight parameters in the jth layer, $j = 1, \ldots, M$. The network weights $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M]$ are learned by applying stochastic gradient descent (SGD), and $\mathcal{F}(\,\cdot\,; \Theta)$ is a nonlinear mapping function which maps $\mathbf{X}$ to the representation of the last hidden layer $\mathbf{Z}$ of the network, i.e.

$$\mathbf{Z} = \mathcal{F}(\mathbf{X}; \Theta). \tag{2.8}$$

Θ is the weight matrix trained by applying backpropagation in the network. The objective function in (2.2) for feature mapping using the hierarchical representation obtained using the neural network is expressed as

$$\mathbf{W}^* = \underset{\mathrm{Tr}(\mathbf{W}^\top \mathbf{Z}\mathbf{C}\mathbf{Z}\mathbf{W})=q}{\arg\min} \mathbf{W}^\top \mathbf{Z}\mathbf{L}\mathbf{Z}^\top \mathbf{W} \tag{2.9}$$

$$= \underset{\mathrm{Tr}(\mathbf{W}^\top \mathbf{W})=q}{\arg\min} \frac{\mathbf{W}^\top \mathbf{Z}\mathbf{L}\mathbf{Z}^\top \mathbf{W}}{\mathbf{W}^\top \mathbf{Z}\mathbf{C}\mathbf{Z}^\top \mathbf{W}}, \tag{2.10}$$

where $\mathbf{C}$ is the Laplacian matrix of the penalty graph $\mathcal{G}^p$. The optimization problems in (2.3), (2.7) and (2.10) can be solved as a generalized eigenvalue problem in the following

$$\mathbf{L}\mathbf{v} = \lambda\, \mathbf{C}\, \mathbf{v}, \tag{2.11}$$

where λ is the set of eigenvalues and $\mathbf{v}$ is the set of eigenvectors.

2.1.2 Dimensionality Reduction

2.1.2.1 Unsupervised Methods

PCA is the classical method for dimensionality reduction by maximizing the variance of data in the projection space. PCA makes three general assumptions:

1. Linearity: PCA is limited to re-expressing the data as a *linear combination* of its basis vectors.

2. The data with large variances contains important structure. Specifically, by assuming the data has a high signal to noise ratio, principal components with larger associated variances represent interesting information.

3. The principal components are orthogonal. It allows an intuitive simplification which makes PCA solvable with linear algebra decomposition techniques.

The detailed process is summarized as follows. Given $\mathbf{X} \in \mathbb{R}^{D \times N}$, where D is the number of dimensions and N is the number of samples, we firstly center the data $\bar{\mathbf{X}} = \mathbf{X} - \frac{1}{N}\mathbf{X}$. Then, the SVD is calculated to the half of the input data which is $\frac{1}{\sqrt{N}}\bar{\mathbf{X}}^\top$, or the eigenvectors of the covariance ($\mathbf{\Sigma_X} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$). $\mathbf{W}$ is a subset of the eigenvectors corresponding to the leading eigenvalues. Finally, we can project $\mathbf{X}$ to the new space $\mathbf{Y}$ with a reduced dimensionality as $\mathbf{Y} = \mathbf{W}^\top\mathbf{X}$.

PCA finds and removes the projection directions with minimal variance, which can be expressed in graph embedding as

$$\mathbf{W}^* = \underset{\mathbf{W}^\top\mathbf{W}=1}{\arg\max}\mathbf{W}^\top\mathbf{\Sigma}\mathbf{W}, \tag{2.12}$$

where $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}(\mathbf{I} - \frac{1}{N}\mathbf{e}\mathbf{e}^\top)\mathbf{X}^\top$ is the covariance matrix, $\mathbf{e}$ is an N-dimensional vector of ones and $\mathbf{I}$ is an identity matrix.

Partial least squares (PLS) regression [20] also finds a linear combination of input basis vectors for regression. It peroforms the eigenanalysis of a variance matrix between inputs $\mathbf{X}$ and $\mathbf{Y}$. As in the case of PCA, the scaling of the variables has a impact on the solutions of the PLS. When expressing the PLS as a graph embedding , the variance between X and $\mathbf{Y}$, i.e.

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{XY}} & \mathbf{0} \end{bmatrix} \tag{2.13}$$

is used in (2.12).

2.1.2.2 Supervised Methods

Linear Discriminant Analysis (LDA) [21] finds a projection by maximizing the ratio of the between-class scatter to the within-class scatter. Let us define by $\boldsymbol{\mu}_c$ the mean vector of the c'th class, formed by N_c samples, and $\boldsymbol{\mu}$ the global mean. Then, LDA optimizes the following criterion:

$$\mathcal{J} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\frac{\mathrm{Tr}(\mathbf{W}^\top\mathbf{P}\,\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top\mathbf{Q}\,\mathbf{W})} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\frac{\mathrm{Tr}(\mathbf{W}^\top\mathbf{X}\mathbf{C}\mathbf{X}^\top\,\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top\mathbf{X}\mathbf{L}\,\mathbf{X}^\top\mathbf{W})}, \tag{2.14}$$

where

$$\mathbf{P} = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top = \mathbf{X}\Big(\sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top - \frac{1}{N}\mathbf{e}\,\mathbf{e}^\top\Big)\mathbf{X}^\top, \tag{2.15}$$

$$\mathbf{Q} = \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top = \mathbf{X}\Big(\mathbf{I} - \sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top\Big)\mathbf{X}^\top. \tag{2.16}$$

Nonlinear extensions with kernels include KDA [22] and KRDA [23].

Locality Preserving Projections (LPP) [24] seeks the k nearest neighbors of the sample $\mathbf{x}_i$, among the samples having the same class label as $\mathbf{x}_i$. It preserves the local information, and obtains a latent space which contains the salient manifold structure. The data projection matrix $\mathbf{W}$ is obtained using the same generalized optimization form as (2.14), while its graph Laplacian matrix $\mathbf{C}$ of the penalty graph is formulated by integrating the discriminative information as follows

$$\mathbf{C} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, & \text{if } \mathbf{x}_i \in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_k(\mathbf{x}_i); \\ 0. & \text{otherwise.} \end{cases} \tag{2.17}$$

2.2 Regularization

A model is trained involving all D dimensions, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. The regularization method also known as shrinkage has the effect of reducing variance and can perform variable selection [25]. It has been largely applied to tackle model overfitting. We discuss several recent techniques for regularization.

2.2.1 ℓ_p-norm Regularization

An additional regularization term is added to the objective function to reduce the model complexity. Suppose we have a loss function $\mathcal{L}(\mathbf{X}, \mathbf{y}|\theta)$, the regularized objective then is

$$\hat{\mathcal{L}}(\mathbf{X}, \mathbf{y}|\theta) = \mathcal{L}(\mathbf{X}, \mathbf{y}|\theta) + \alpha R(\theta), \tag{2.18}$$

where $R(\theta)$ is the regularization term, and α is a control parameter.

The general form of ℓ_p-norm based regularization is $R(\theta) = \sum_j \|\theta_j\|_p^p$. When $p \leq 1$, the objective is a convex optimization problem. In particular, the ℓ_2-norm regularization is commonly used which is known as weight decay. When $p \leq 1$, the resulting regularization exploits the sparsity of the objective function with a non-convex optimization.

2.2.2 Dropout and DropConnect

Dropout [26, 27] has been originally proposed as a regularization strategy for neural networks training. It prunes the neurons in the network to effectively regularize the model in an online fashion. It can be also considered analogously as a bagging ensembles of many large neural networks, which learns the network output weights. The outputs of the synthetic hidden layer by dropout is written as

$$\mathbf{z} = \mathbf{m}_{i,t} \circ \boldsymbol{\psi}_{i,t}(\mathbf{W}^\top \mathbf{x}), \quad i = 1, \cdots, N, \ t = 1, \cdots, T, \tag{2.19}$$

where $\circ$ denotes the Hadamard product of two vectors and $\mathbf{m}_{i,t} \in \mathbb{R}^{n_l}$ is a binary mask vector with each element equal to 1 with probability p and equal to 0 with $(1-p)$, and n_l the number of neurons in the lth layer. We denote i as the number of samples, and t is the epoch of network training. The binary mask $m_{i,t}$ is selected independently for each sample from a Bernoulli distirbution and changes over the training iterations, and therefore, the trained model is a bagged ensemble of neural networks. The difference with a traditional bagging method is that the neural networks using bagging share parameters from the original (full) neural network.

Besides dropout regularization, we can also set the elements of output weight matrix $\mathbf{W}$ to zero, which effectively drops the connections between neurons. The synthetic hidden layer outputs by DropConnect are

$$\mathbf{z} = \psi_{i,t}(\mathbf{M} \circ \mathbf{W}^\top)\mathbf{x}), \quad t = 1, \cdots, N, \; t = 1, \cdots, T, \tag{2.20}$$

where $\mathbf{M}_{i,t} \in \mathbb{R}^{n_l \times n_{(l-1)}}$ is a binary mask matrix with its elements equal to 1 with probability p and $(1-p)$ otherwise. Both Dropout and DropConnect use the masked versions of weight for nerual network training.

2.3 Multi-view learning

2.3.1 Overview

In general, a view is referred to as a group of features extracted from a domain or modality. The modern process of data acquisition across various sensory modalities gives rise to the heterogeneous property of data. Multiple features can be generated in diverse domains or using different descriptors to represent the same data sample. Multi-view learning is the set of methods which leverage the information of heterogeneous data [4]. The goal of multi-view learning is therefore to integrate multiple views to make effective decisions.

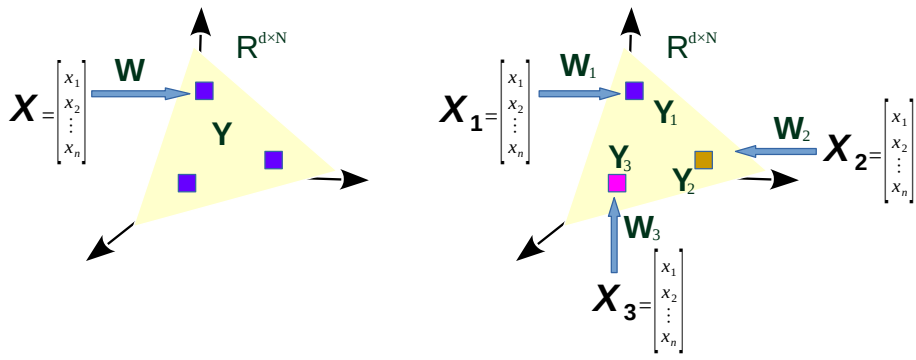
There is a considerable difference between conventional machine learning algorithms and multi-view learning. The former takes a single-view input or a concatenation of multiple views, trains a model and generates an output of single view. By contrast, its multi-view counterpart joinly learns a model by optimizing a function from the multiple views, and make predictions using the enriched information. Methods to exploit the sensory redundancies, which contains both common and complementary information, can be classified into three groups, including subspace learning, co-training and multiple kernel learning.

We can categorize the methods of multi-view learning based on the tasks that we want to achieve [28]: representation, translation, alignment, fusion and co-learning. Representation learning is a task to extract a feature representation by learning a mapping

function from the original data. Subspace learning providing a compact representation belongs to this category. Translation considers feature mapping from one modality to another, so that a cross-view learning can be achieved. Alignment is to find the matches in elements between modalities. Fusion aims to integrate the multi-view information for making predictions. Multiple kernel learning belongs to this category. Finally, co-learning considers transferring knowledge across modaltiies. Co-training and zero-shot learning are examples of this type of methods. We will elborated the important multi-view learning methods in the following sections.

2.3.2 Multi-view Subspace Learning

The challenge in multi-view learning is that there exists a large discrepency between views. Mapping each of the views to their own subspace does not provide good matches between projected features. Multi-view subspace learning mitigates the problem by projecting them into a common latent space by optimizing a joint criterion. We present this idea in Figure 2.1.



Single-view subspace learning Multi-view subspace learning

Figure 2.1: Single-view subspace learning projects one input to a latent space, while multi-view subspace learning leverages a common space from multiple inputs.

2.3.3 Unsupervised Multi-view Subspace Learning Methods

2.3.3.1 Linear CCA

Canonical Correlation Analysis (CCA) [29, 30] is a conventional statistical technique which finds the maximum correlation between two sets of data samples $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times N}$ using the linear combination $\mathbf{Y}_1 = \mathbf{W}_1^\top \mathbf{X}_1$ and $\mathbf{Y}_2 = \mathbf{W}_2^\top \mathbf{X}_2$. $\mathbf{W}_1$ and

$\mathbf{W}_2$ are determined by optimizing:

$$\mathcal{J} = \arg\max_{\mathbf{W}_1,\mathbf{W}_2} \text{corr}(\mathbf{W}_1^\top \mathbf{X}_1, \mathbf{W}_2^\top \mathbf{X}_2) \tag{2.21}$$

$$= \arg\max_{\mathbf{W}_1,\mathbf{W}_2} \frac{\mathbf{W}_1^\top \mathbf{\Sigma}_{12} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top \mathbf{\Sigma}_{11} \mathbf{W}_1} \cdot \sqrt{\mathbf{W}_2^\top \mathbf{\Sigma}_{22} \mathbf{W}_2}}, \tag{2.22}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \bar{\mathbf{X}}_1 \bar{\mathbf{X}}_1^\top & \bar{\mathbf{X}}_1 \bar{\mathbf{X}}_2^\top \\ \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_1^\top & \bar{\mathbf{X}}_2 \bar{\mathbf{X}}_2^\top \end{bmatrix} \tag{2.23}$$

2.3.3.2 Kernel CCA

Kernel CCA finds the maximum correlation between two views after mapping them to the kernel space [31]. It is expressed mathematically as

$$\mathcal{J} = \arg\max_{\mathbf{W}_1,\mathbf{W}_2} \text{corr}(\mathbf{W}_1^\top \mathbf{\Phi}_1, \mathbf{W}_2^\top \mathbf{\Phi}_2) \tag{2.24}$$

Using the kernel trick [17] and the Representer Theorem in (2.5), the objective function for the kernel CCA becomes

$$\mathcal{J} = \arg\max_{\text{Tr}(\mathbf{A}_1^\top \mathbf{C} \mathbf{A}_2) = p} \frac{\mathbf{A}_1^\top \mathbf{K}_1 \mathbf{K}_2 \, \mathbf{A}_2}{\sqrt{\mathbf{A}_1^\top \mathbf{K}_1 \mathbf{K}_1 \, \mathbf{A}_1} \cdot \sqrt{\mathbf{A}_2^\top \mathbf{K}_2 \mathbf{K}_2 \, \mathbf{A}_2}}. \tag{2.25}$$

2.3.3.3 Deep CCA

Deep CCA maximizes the correlation between a pair of views by learning nonlinear representations from the input data through multiple stacked layers of neurons [32, 33]. A linear CCA layer is added on top of both networks, and the inputs to the CCA layer depend on the network outputs $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Similar to the nonlinear case in (2.25), a modified objective function $\min_{\mathbf{W}_1,\mathbf{W}_2} -\frac{1}{N} \text{Tr}\left(\mathbf{W}_1{}^\top \mathbf{Z}_1 \, \mathbf{Z}^\top \mathbf{W}_2\right)$ is optimized, where $\mathbf{W}_1, \mathbf{W}_2$ are the projection matrices in the CCA layer, and the correlated outputs are $\mathbf{Y}_1 = \mathbf{W}_1^\top \mathbf{Z}_1$ and $\mathbf{Y}_2 = \mathbf{W}_2^\top \mathbf{Z}_2$. A modified SGD method is developed with respect to the inputs $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to the linear layer, which are also the outputs from the two networks. The objective function is expressed as $\text{Tr}\left(\mathbf{W}_1^\top \mathbf{Z}_1 \, \mathbf{Z}_2^\top \mathbf{W}_2\right) = \text{Tr}(\mathbf{T}^\top \mathbf{T})^{\frac{1}{2}}$, which describes the correlation as the sum of the top d singular vectors of $\mathbf{T} = \mathbf{\Sigma}_{11}^{-1/2} \mathbf{\Sigma}_{12} \mathbf{\Sigma}_{22}^{-1/2}$ whose definition can be found in [34]. The projection matrices are obtained from the singular value decomposition of $\mathbf{T}$, as $\mathbf{T} = \mathbf{W}_1 \mathbf{D} \mathbf{W}_2^\top$. The gradient is then computed as

$$\frac{\partial(\text{Tr}(\mathbf{T}^\top \mathbf{T})^{\frac{1}{2}})}{\partial \mathbf{Z}_1} = 2\Delta_{11} \mathbf{Z}_1 + \Delta_{12} \mathbf{Z}_2, \tag{2.26}$$

where

$$\Delta_{12} = \mathbf{\Sigma}_{11}^{-1/2} \mathbf{W}_1 \mathbf{W}_2^\top \mathbf{\Sigma}_{22}^{-1/2} \tag{2.27}$$

$$\Delta_{11} = -\frac{1}{2} \mathbf{\Sigma}_{11}^{-1/2} \mathbf{W}_1 \mathbf{D} \mathbf{W}_1^\top \mathbf{\Sigma}_{11}^{-1/2}. \tag{2.28}$$

and $\partial(\text{Tr}(\mathbf{T}^\top \mathbf{T})^{\frac{1}{2}})/\partial \mathbf{Z}_2$ takes the symmetric form of the above definition.

2.3.4 Supervised Multi-view Subspace Learning Methods

2.3.4.1 Multi-view Discriminant Analysis (MvDA)

MvDA [35] is the multi-view verison of LDA which maximizes the ratio of the traces of the between-class scatter matrix to that of the within-class scatter matrix. Its objective function is

$$\mathcal{J} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{S}_B)}{\mathrm{Tr}(\mathbf{S}_W)}, \tag{2.29}$$

where the between-class scatter matrix is

$$\mathbf{S}_B = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \left(\sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top - \frac{1}{N} \mathbf{e}\,\mathbf{e}^\top \right) \mathbf{X}_j^\top \mathbf{W}_j, \tag{2.30}$$

and the within-class scatter matrix is

$$\mathbf{S}_W = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \left(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top \right) \mathbf{X}_j^\top \mathbf{W}_j. \tag{2.31}$$

$\mathbf{W}$ contains the eigenvectors of the matrix $\mathbf{S} = {\mathbf{S}_W}^{-1}\mathbf{S}_B$ corresponding to the leading d eigenvalues λ_i.

2.3.5 Multi-view Uncorrelated Discriminant Analysis (MvUDA)

Multi-view Uncorrelated Discriminant Analysis is an extension of MvDA inspired by the uncorrelated discriminant transform in [36]. We consider the method in 2 views. Following the same form of objective function as MvDA in (2.29), it multiplies a uncorrelated term with between-class scatter matrix, and the solution is

$$\begin{bmatrix} \mathbf{P_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{b_1} & \gamma\boldsymbol{\Sigma}_{12} \\ \gamma\boldsymbol{\Sigma}_{12} & \boldsymbol{S}_{b_2} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{S}_{w_1} & \mathbf{0}_{12} \\ \mathbf{0} & \sigma\boldsymbol{S}_{w_2} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix}, \tag{2.32}$$

where $\mathbf{P}_1$ and $\mathbf{P}_2$ uncorrelate the between-class scatters, and are expressed by

$$\mathbf{P}_1 = \mathbf{I} - \boldsymbol{S}_{w_1}\mathbf{W}_1^\top(\mathbf{W}_1\boldsymbol{S}_{w_1}\mathbf{W}_1^\top)^{-1}\mathbf{W}_1,$$
$$\mathbf{P}_2 = \mathbf{I} - \boldsymbol{S}_{w_2}\mathbf{W}_2^\top(\mathbf{W}_2\boldsymbol{S}_{w_2}\mathbf{W}_2^\top)^{-1}\mathbf{W}_2.$$

γ and σ are scaling parameters.

2.3.6 Semi-supervised Learning

Co-training method [37] is a semi-supervised learning method which maximizes the mutual agreement on a pair of distinct views of the unlabeled data. It provides a superior classification performance under the following assumptions.

- The training samples contain two sufficient sets of features $(\mathbf{X}_1, \mathbf{X}_2)$, while each sample has two corresponding views $(\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2])$.

- The two views are independent given the class label, i.e.

$$P(\mathbf{X}_1|\mathbf{X}_2, \mathbf{y}) = P(\mathbf{X}_1|\mathbf{y}), \tag{2.33}$$
$$P(\mathbf{X}_2|\mathbf{X}_1, \mathbf{y}) = P(\mathbf{X}_2|\mathbf{y}).$$

- The two views are consistent:

$$\exists f_1, f_2 : f_{\text{co-trained}}(\mathbf{X}) = f_1(\mathbf{X}_1) = f_2(\mathbf{X}_2). \tag{2.34}$$

The method has been well recognized, and many efforts have been made beyond its original usage in text mining for search engine. Expectation-maximization (EM) was successfully applied to classify new samples between classifiers using a probabilistic approach [38]. Multi-view spectral clustering is enabled by introducing a co-regularization to the clustering in [39]. Bayesian view of co-training is developed in [40]. Co-training is also extended for learning to rank in [41].

2.4 Applications

Multi-view subspace learning can be used in numerous application domains. In the following, we briefly describe the major applications used in the thesis to evaluated the performance of the proposed methods.

2.4.1 Cross-modal Multimedia Retrieval

Traditional multimedia retrieval applications consider unimodal scenarios. Typical examples in this case include text search by matching strings or related topics, and content-based image retrieval [42, 43]. In contrast to unimodal solutions, multimodal retrieval systems have been developed mainly about image retrieval using text queries [44, 45, 46, 47]. Large multimedia repositories such as TRECVID [48] and ImageCLEF [49] have been collected to study and evaluate the retrieval over multiple modalities. However, these methods are still based on the unimodal approach, e.g. people use text queries to match the tags surrounding the images to search for the relevant images.

There is an increasing amount of efforts in cross-modal retrieval. Traditionally, the text annotations of images can be scarsely found. Thanks to sensor technologies and popularity of internet services, there is an explosion of multimedia content avaiable online, and these data are richly annotated with full descriptions. Manifold learning has been successfully applied from a matrix of distances between multimodal objects [50, 51, 52]. The multimodal distances are formulated as a function of distances between each pair of modalities, which allows mis-matched pairs. However, it limits the queries to the training set which is used to learn the manifold. Recently, a new cross-modal retrieval method [3] is developed by exploiting the correlation between modalities and mapping the feature

in the latent space towards to the semantic labels. Superior retrieval performance has demonstrated the effective combination of correlation and semantic matching.

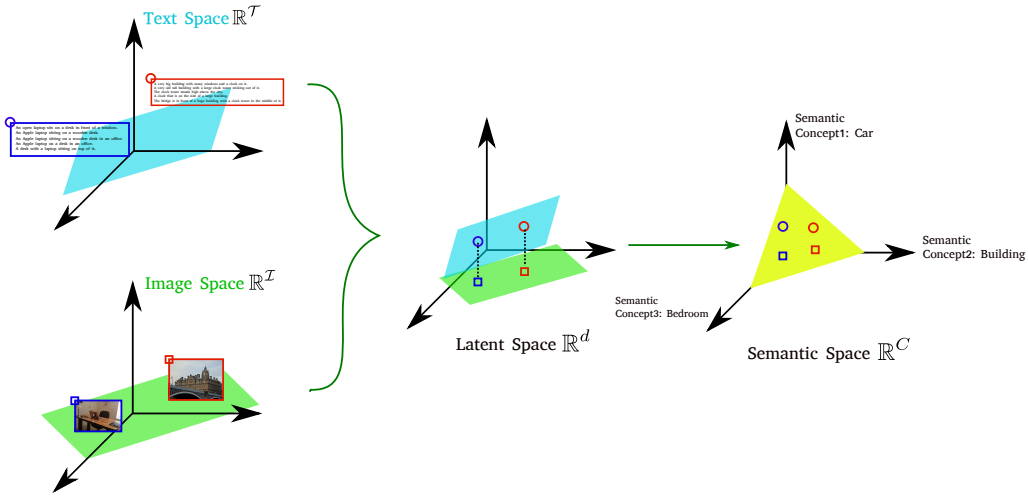


Figure 2.2: An illustration of cross-modal image retrieval.

2.4.2 Zero-shot Object Recognition

Zero-shot object recogntion [53] is an emerging topic which aims to recognize objects of unseen classes. It is inpired from the real-world scenario of human categorizing new objects or generalizing novel concepts. The approach has a strong connection with learning to learn [54], and lifelong learning [55]. The main idea is to establish a relation from the objects in the source domain (seen classes) to the target domain (unseen classes) using a universal semantic representation. There are several ways to generate the semantic representation or attributes, which includes user-defined attributes [56, 57], relative attributes [58, 59, 7], and data-driven attributes [60, 61]. The transferrable knowledge enables the object recognition of unseen classes. Visual features are mapped to the latent space of semantic representations. The unlabeled target class is projected in the same space. One major problem of zero-shot recognition is that data distribution of the source classes and target classes is different, and therefore a domain shift is found when projecting both data from both domains to the same latent space. The problem is allieviated by introducing a multi-view semantic latent space which fuses data from multiple modalities [7].

2.4.3 Face-sketch Recognition

Face-sketch recognition is an important application enabling searches of potential suspects in a mugshot database created by law enforcement. The main idea is to shortlist the photos in the database which may match to the face of the suspect. Usually, sketches are

drawn based on the description by the eyewitness, and the most distinctive facial features are presented on the sketches. However, sketching a face involves many psychological factors, which may result in misleading face recognition. Some examplar images from the face-sketch database are shown in Figure 2.3.

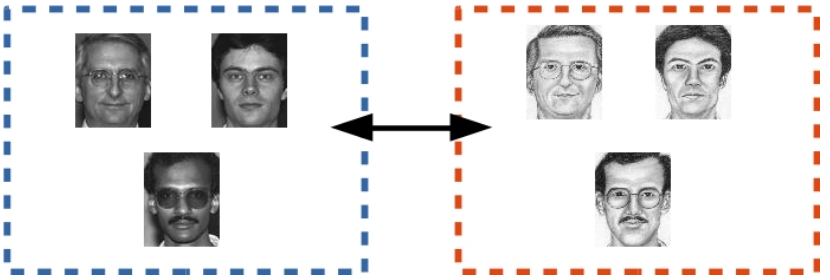


Figure 2.3: Examplary face-sketch image pairs in the CUFSF dataset [1].

Subspace learning has been successfully applied in the face-sketch recognition [62, 63]. Linear mappings between faces and sketches using PCA was introduced in [62]. Kernel-based LDA was also applied as the nonlinear mappings based on image patches [63]. Another extension in [64] is to perform random feature sampling and calculate kernel prototype similarities before applying LDA. There are many efforts which project both modalities into a common subspace. For example, coupled discriminant analysis is introduced in [65] to learn from faces and sketches in a common latent space. Directly learning the image filters from the raw faces and sketches simultaneously also shows its effectiveness in heterogeneous face recognition [66]. Image patches are represented using Markov random fields to incorporate the spatial information in patch neighborhoods [67]. A novel similarity metric is also proposed to calculate the distance between faces and sketches.

2.4.4 Learning to Rank

Ranking problems can be found in numerous applications, for example ratings of food or movies, image retrieval and ranking [68, 69], image quality ratings [70], online advertising [71], and text summarization [41]. In general, there is a series of data pre-processing and indexing to generate the pairs of queries and samples for matching. The ranker, which is the key comoponent, provides a relevance score between each pair of query and sample. The score can be calculated based on some heuristic measure or learning approach.

Learning to rank aims to optimize the combination of data representation for ranking problems [72]. We present its framework in Figure 2.4. Suppose we have N training samples, which consists of $\mathbf{q}^{tr}, \mathbf{X}^{tr}$ and $\mathbf{y}^{tr}$. $\mathbf{X}^{tr} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^\top$ is the feature vectors of the training data, where $\mathbf{x}_n \in \mathbb{R}^D$ represents its nth sample. $\mathbf{q}^{tr} = [q_1, q_2, \cdots, q_n]$ is the corresponding query and $\mathbf{y}^{tr} = [y_1, y_2, \cdots, y_n]$ is the relevance, respectively. A

ranking model h is learned over the training data, and makes predictions over the test data $\mathbf{q}^{te}, \mathbf{X}^{te}$. We predict its relevance $h(\mathbf{q}^{te}, \mathbf{X}^{te})$ using the trained model h.

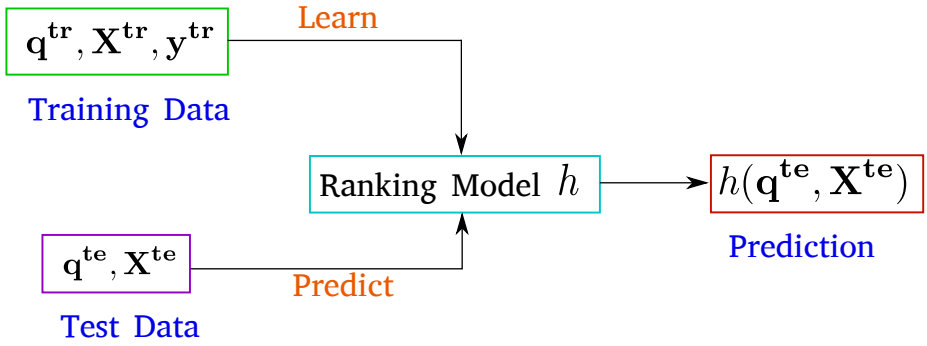


Figure 2.4: The framework of Learning to Rank.

Solutions to this problem can be decomposed into several key components, including the input feature, the output vector and the scoring function. The framework is developed by training the scoring function from the input feature to the output ranking list, and then, scoring the ranking of new data. Traditional methods also include engineering the feature using the PageRank model [73], for example, to optimally combine them for obtaining the output. Later, research was focused on discriminatively training the scoring function to improve the ranking outputs. The ranking methods can be classified in three categories for the scoring function: the pointwise approach, the pairwise apporach, and the listwise approach.

The pairwise approach is considered in the thesis and therefore reviewed thoroughly as follows. A neural network which learns a preference function was developed in [74] to directly evaluated the pairwise order between pairs of documents. RankNet [75] learns a neural network to optimize the pairwise ranking loss using a cross-entropy loss. The pairwise methods generally assume the scoring function to be linear [76], so the ranking data can be easily transformed to orders in pairs. The transformed data enables a binary classification for ranking, and therefore numerous classifiers have been applied. Adaboost algorithm [77] was successfully applied by iteratively reducing the classfication errors between each pair of documents, which can subsequently improve the overall output. Ranking SVM [78] adopted SVM to perform pairwise classification. GBRank used Gradient Boost Trees [79] in ranking documents. Semi-supervised multi-view ranking (SmVR) [41] is a co-training extension to ranking. Recently, there is an increasing amount of research in optimizing the evaluation metric for ranking. Examples include AdaRank [80], which optimizes the ranking errors iteratively, and LambdaRank [81]. While certain success has been obtained by the aforementioned methods, ranking multi-facet documents is important yet few can be found in literature [82, 83, 84].

2.4.4.1 Bipartite Ranking

The pairwise transform is critical for the success in ranking and therefore described explicitly in this section. The training data is defined in query-sample pairs $\{(\mathbf{x}_i^q, \mathbf{y}_i^q)\}$, where $q \in \{1, 2, \ldots, Q\}$, $\mathbf{x}_i^q \in \mathrm{R}^d$ is the d-dimensional feature vector for the pair of query q, the i-th sample, $\mathbf{y}_i^q \in \{0, 1\}$ is the relevance score, and the number of query-specific samples is N_q. The pairwise transformation to generate the query-sample pairs, so that only the samples that belong to the same query are evaluated [76].

The relevance between each pair is defined as

$$\mathbf{p}_i^q(\phi) = \frac{1}{1 + \exp(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_q))},$$

where $\phi : \mathbf{x} \to \mathbb{R}$ is the linear scoring function as $\phi(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$, which maps the input feature vectors to the scores. Due to its linearity, we can transform the feature vectors and relevance score into $(\mathbf{x}'_k, \mathbf{y}'_k) = (\mathbf{x}_q - \mathbf{x}_i, \mathbf{y}_i^q)$. In case of the ordered list $(\mathbf{r})$ as the raw input, each data sample $\mathbf{x}_i$ paired with its query $\mathbf{x}_q$ is investigated, and their raw orders $(\mathbf{r}_i, \mathbf{r}_q)$ are transformed as $\mathbf{y}_i^q = 1$, if $\mathbf{r}_i < \mathbf{r}_q$; $\mathbf{y}_i^q = 0$, else if $\mathbf{r}_i > \mathbf{r}_q$. In pairwise ranking, the relevance $\mathbf{y}_i^q = 1$, if the query and sample are relevant, and $\mathbf{y}_i^q = 0$, otherwise.

The feature difference $(\mathbf{x}'_k, \mathbf{y}'_k)$ becomes the new feature vector as the input data for nonlinear transforms and subspace learning. Therefore, the probability can be rewritten as

$$\mathbf{p}_k(\phi) = \frac{1}{1 + \exp(-\phi(\mathbf{x}'_k))} = \frac{1}{1 + \exp(-\mathbf{a}^\top \mathbf{x}'_k)}. \tag{2.35}$$

The ranking loss is formulated as the cross entropy loss such that,

$$\begin{aligned} \ell_{\text{Rank}} &= \arg\min \sum_{q=1}^{Q} \sum_{i=1}^{N_q} \big(\mathbf{y}_i^q \log \mathbf{p}_i^q) + (1 - \mathbf{y}_i^q) \log \mathbf{p}_i^q)\big) \\ &= \arg\min \sum_{k=1}^{K} \big(\mathbf{y}'_k \log \mathbf{p}_k) + (1 - \mathbf{y}'_k) \log \mathbf{p}_k)\big), \end{aligned} \tag{2.36}$$

which is proved in [75] that it is an upper bound of the pairwise 0-1 loss function and optimized using gradient descent. The logistic regression or softmax function in neural networks can be used to learn the scoring function.

3 Contributions

This chapter describes the novel contributions of the thesis. We will begin with the generalized multi-view embedding method, which is the major contribution of the thesis. Its extention to a multi-view non-parametric method exploiting the class boundary structure and discrepancy in views is subsquently described. Additionally, the dropout regularization is introduced to the linear multi-view analysis. Finally, we will present composite ranking methods for ranking problems which enhance the joint ranking with minimum loss from each ranking source.

3.1 Generalized Multi-view Embedding

We propose a unified solution for multi-view subspace learning which generalizes several statistical, supervised and nonlinear embeddings. Here, we solve the generalized optimization problem

$$\mathcal{J} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{P} \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{Q} \mathbf{W})} \tag{3.1}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the matrices describing properties of the data to be maximized and minimized, respectively, through embedding. We consider it as the uniform objective function, reaching out to a large number of subspace learning methods. A generalized eigenvalue problem is addressed when maximizing the criterion:

$$\mathbf{P}\mathbf{W} = \boldsymbol{\rho}\, \mathbf{Q}\mathbf{W}, \tag{3.2}$$

and the solution is given in the following form:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_V \end{pmatrix} \text{ and } \boldsymbol{\rho} = \sum_{i=1}^{d} \lambda_i. \tag{3.3}$$

$\mathbf{W}, \rho$ are the generalized eigenvector and the sum of the top d generalized eigenvalues λ_i, respectively. $\mathbf{W}$ contains the projection matrices of all views, and $\boldsymbol{\rho}$ is the value of Rayleigh quotient in (3.1). The nonlinear multi-view embeddings can be achieved by

using kernel-based mappings, or (deep) neural networks optimized by SGD. In the case of linear projections, i.e. when $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$, we derive the objective function based on [15, 85] as follows

$$\mathcal{J} = \underset{\mathbf{W}^\top \mathbf{W} = \mathbf{I}}{\arg\max} \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L}' \mathbf{X}^\top \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L} \mathbf{X}^\top \mathbf{W})}. \tag{3.4}$$

In the kernel case, we have

$$\mathcal{J} = \underset{\mathbf{A}^\top \mathbf{K} \mathbf{A} = \mathbf{I}}{\arg\max} \frac{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}'\, \mathbf{K} \mathbf{A})}{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}\, \mathbf{K} \mathbf{A})}. \tag{3.5}$$

In the above, we define $\mathbf{L}$ as the total graph Laplacian matrix. Similarly, the penalty graph Laplacian matrix is denoted by $\mathbf{L}'$.

In the case of the nonlinear embedding using neural networks, we apply a joint linear embedding layer on top of the neural networks $\mathcal{F}_v$, where $v = 1, \cdots, V$. The scheme is presented in Figure 3.1. We train V sub-networks whose outputs are projected to a common subspace using a linear projection $\mathbf{W}_v$. We denote $\mathcal{F}(\mathbf{X}) = [\mathcal{F}_1(\mathbf{X}_1), \cdots, \mathcal{F}_V(\mathbf{X}_V)]^\top$ as the concatenation of the neural network outputs. By doing so, the objective has the same form as in the linear case. By following the direction of the gradient for training the neural network, we optimize the Rayleigh quotient criterion with respect to the nonlinear feature representation from each view in the last hidden layer of the networks. The entire network is trained in a single optimization process.

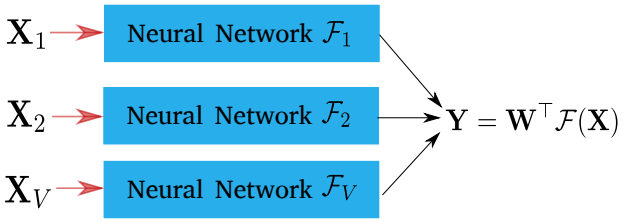


Figure 3.1: Schematic presentation of Multi-view (Deep) Embedding Networks.

We illustrate the proposed framework graphically in Figure 3.2. Suppose we initially extract three types of low-level features from images, texts, and intermediate representations. The multimodal features are projected using either linear or nonlinear projections to the common latent space. The projected features characterize the properties of the intra-view compactness and inter-view separability based on the Rayleigh quotient criterion.

3.1.1 Scaling Up the Inter-view and Intra-view Covariance Matrices

Numerous statisitical subspace learning methods can be generalized in the form of (3.1) by scaling up the inter-view and intra-view covariance matrices. Multi-view CCA (MvCCA) presented in [P1] maximizes the correlation between all pairs of views. Its objective can be rephrased as maximizing the inter-view covariance while minimizing the intra-view

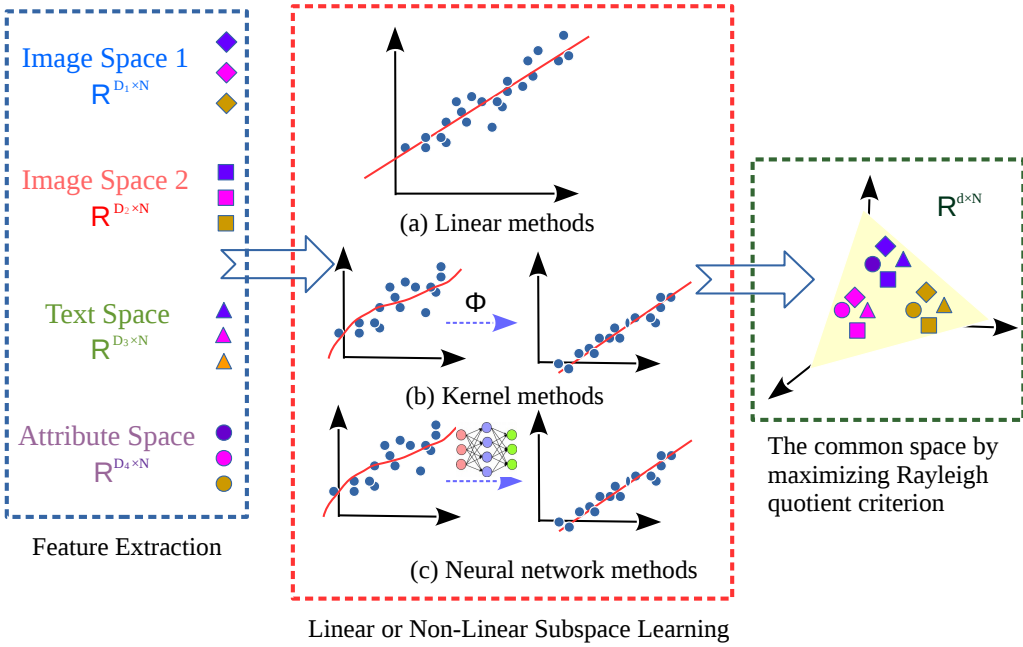


Figure 3.2: Overview of the generalized multi-view embedding: Features from different modalities are extracted and either linearly or nonlinearly mapped into the common subspace by maximizing the Rayleigh quotient criterion [P1].

covariance in the latent space. Therefore, we consider inter-view covariance matrices between different view representations in $\mathbf{P}$ and the covariance matrices of each view in $\mathbf{Q}$. Multi-view PLS (MvPLS) maximizes the inter-view covariance directly, and its difference with MvCCA is the intra-view minimization. Taking the class discrimination into consideration, the proposed multi-view modular discriminant analysis (MvMDA) extends to separate the data of different classes between views while making the intra-class data compact. We present the structure of $\mathbf{P}$ and $\mathbf{Q}$ for each method in Table 3.1.

3.1.2 Linear Subspace Learning

When the subspace projection is linear, we can obtain the latent feature vectors from each view as

$$\mathbf{Y}_v = \mathbf{W}_v^\top \mathbf{X}_v, \tag{3.6}$$

which corresponds to the case on the top of Figure 3.2 using the linear feature mappings. Its projection matrix is obtained by directly solving the generalized eigenvalue problem in (3.2). Multi-view CCA minimizes the diagnal matrix and maximizes the off-diagonal matrix of the total covariance matrix shown in Table 3.1. we derive its projection matrix

Table 3.1: The matrices $\mathbf{P}$ and $\mathbf{Q}$ for the proposed multi-view CCA, PLS and MvMDA [P1].

	$\mathbf{P}$	$\mathbf{Q}$
MvCCA	$\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1V} \\ \mathbf{\Sigma}_{21} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{V1} & \mathbf{\Sigma}_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{VV} \end{bmatrix}$
MvPLS	$\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1V} \\ \mathbf{\Sigma}_{21} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{V1} & \mathbf{\Sigma}_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$
MvMDA	$\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1V} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{V1} & \mathbf{P}_{V2} & \cdots & \mathbf{P}_{VV} \end{bmatrix}$	$\begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{VV} \end{bmatrix}$

by optimizing the criterion

$$\mathcal{J} = \underset{\mathbf{W}_v, v=1,\ldots,V}{\arg\max} \frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_i^\top \mathbf{W}_i\right)}, \tag{3.7}$$

where the Laplacian matrix $\mathbf{L} = \mathbf{I} - \frac{1}{N}\mathbf{e}\,\mathbf{e}^\top$.

Multi-view PLS considers the penalty graph only, and its objective is to maximize the cross-covariance matrices between different views, as follows:

$$\mathcal{J} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\, \mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j\right). \tag{3.8}$$

We propose two new methods for multi-view LDA. The first approach is the multi-view extension of the standard LDA, and maximizes the distance between class centers of each view pair. Its between-class scatter $\mathbf{S}_B$ is

$$\begin{aligned} \mathbf{S}_B &= \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{p=1}^{C}\sum_{\substack{q=1 \\ p\neq q}}^{C} (\boldsymbol{m}_p^i - \boldsymbol{m}_q^j)(\boldsymbol{m}_p^i - \boldsymbol{m}_q^j)^\top \\ &= \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_B \mathbf{X}_j^\top \mathbf{W}_j, \end{aligned} \tag{3.9}$$

and the between-class Laplacian matrix is

$$\mathbf{L}_B = \begin{cases} 2\sum_{p=1}^{C}\sum_{\substack{q=1\\p\neq q}}^{C}\left(\frac{V}{N_p^2}\mathbf{e}_p\,\mathbf{e}_p^\top - \frac{1}{N_pN_q}\mathbf{e}_p\,\mathbf{e}_q^\top\right) & \text{if } i=j, \\ -2\sum_{p=1}^{C}\sum_{\substack{q=1\\p\neq q}}^{C}\frac{1}{N_pN_q}\mathbf{e}_p\,\mathbf{e}_q^\top & \text{if } i\neq j. \end{cases} \tag{3.10}$$

$\boldsymbol{m}_p^i$ denotes the mean from the ith view of the pth class in the latent space, and $\mathbf{e}_p$ is the N-dimensional class vector, with N_p as the number of samples in the pth class. The class q is different from the class p.

Moreover, we also consider maximizing the distance between different view-specific class centers in the between-class scatter matrix. As the objective is to maximize the sample distances from the subclasses of each specific view, we name the method as Multi-view Modular Discriminant Analysis (MvMDA). The corresponding multi-view between-class scatter matrix is

$$\begin{aligned}\mathbf{S}'_B &= \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{p=1}^{C}\sum_{\substack{q=1\\p\neq q}}^{C}(\boldsymbol{m}_p^i-\boldsymbol{m}_q^i)(\boldsymbol{m}_p^j-\boldsymbol{m}_q^j)^\top \\ &= \sum_{i=1}^{V}\sum_{j=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{L}'_B\mathbf{X}_j^\top\mathbf{W}_j,\end{aligned} \tag{3.11}$$

and the Laplacian matrix is

$$\mathbf{L}'_B = 2\sum_{p=1}^{C}\sum_{q=1}^{C}\left(\frac{1}{N_p^2}\mathbf{e}_p\,\mathbf{e}_p^\top - \frac{1}{N_pN_q}\mathbf{e}_p\,\mathbf{e}_q^\top\right). \tag{3.12}$$

[P1] also provides a detailed derivation of the difference between the two multi-view LDA methods, which is that $\mathbf{S}_B$ has the term $\frac{1}{N_c^2}(V-1)\sum_{i=1}^{V}\sum_{c=1}^{C}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{e}_c\mathbf{e}_c^\top\mathbf{X}_i^\top\mathbf{W}_i$, while $\mathbf{S}'_B$ has the term $\frac{1}{N_c^2}\sum_{i=1}^{V}\sum_{\substack{j=1\\j\neq i}}^{V}\sum_{c=1}^{C}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{e}_c\mathbf{e}_c^\top\mathbf{X}_j^\top\mathbf{W}_j$. This difference suggests that the first proposal only considers the maximum of the intra-view distances, while the second proposal can maximize the distance between different views. It was also shown in the experiments that the second approach achieves better results. The within-class scatter matrix for both methods is formulated by directly scaling the single-view scatter matrix, i.e.

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\left(\mathbf{I}-\sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top\right)\mathbf{X}_i^\top\mathbf{W}_i \\ &= \sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Q}_{ii}\,\mathbf{W}_i,\end{aligned} \tag{3.13}$$

where $\mathbf{Q}_{ii} = \mathbf{X}_i \mathbf{L}_W \mathbf{X}_i^\top$, and $\mathbf{L}_W = \mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top$. From (3.9) and (3.13), it is shown that the between-class and within-class scatters are equivalent to the projected inter-view and intra-view covariance, respectively. That is, the objective function is optimized as follows

$$\mathcal{J} = \underset{\mathbf{W}_v, v=1,\ldots,V}{\arg\max} \frac{\mathrm{Tr}\big(\sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_B^* \mathbf{X}_j^\top \mathbf{W}_j\big)}{\mathrm{Tr}\big(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_W \mathbf{X}_i^\top \mathbf{W}_i\big)}, \tag{3.14}$$

where $\mathbf{L}_B^*$ is denoted as the Laplacian matrix of either $\mathbf{L}_B$ or $\mathbf{L}_B'$.

3.1.3 Kernel-based Nonlinear Subspace Learning

We also derive the nonlinear multi-view embedding using kernel-based feature mappings. Exploiting the kernel trick in (3.30) and the Representer theorem in (2.5), the mapping is expressed as follows

$$\mathbf{Y}_v = \mathbf{A}_v^\top \mathbf{\Phi}_v^\top \mathbf{\Phi}_v = \mathbf{A}_v^\top \mathbf{K}_v. \tag{3.15}$$

The criterion of kernel multi-view CCA is then,

$$\mathcal{J} = \underset{\mathbf{K}_v, v=1,\ldots,V}{\arg\max} \frac{\mathrm{Tr}\big(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_j \mathbf{A}_j\big)}{\mathrm{Tr}\big(\sum_{i=1}^{V} \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_i \mathbf{A}_i\big)}, \tag{3.16}$$

where the matrix $\mathbf{A}_v$ can be obtained from (3.2).

Kernel multi-view PLS maximizes the covariance between different view pairs in the kernel space and its objective function is

$$\mathcal{J} = \underset{K_v, v=1,\ldots,V}{\arg\max} \mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L} \mathbf{K}_j \mathbf{A}_j\right). \tag{3.17}$$

The criterion for kernel multi-view discriminant analysis is

$$\mathcal{J} = \underset{K_v, v=1,\ldots,V}{\arg\max} \frac{\mathrm{Tr}\big(\sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L}_B^* \mathbf{K}_j \mathbf{A}_j\big)}{\mathrm{Tr}\big(\sum_{i=1}^{V} \mathbf{A}_i^\top \mathbf{K}_i \mathbf{L}_W \mathbf{K}_i \mathbf{A}_i\big)} \tag{3.18}$$

3.1.4 Nonlinear Subspace Learning using (Deep) Neural Networks

Moreover, neural networks are employed for multi-view embedding, and its nonlinear data projection of each view through feature mappings is

$$\mathbf{Y}_v = \mathbf{W}_v^\top h(\mathbf{X}_v; \mathbf{B}_v) = \mathbf{W}_v^\top \mathcal{H}_v. \tag{3.19}$$

Since the neural network outputs $\mathcal{H}_v, v = 1, \cdots, V$ are combined by a linear layer as shown in Figure 3.1, the assembled networks are jointly optimized and the parameters $\mathbf{B}_v$ of all networks are optimized together accordingly. In each training epoch, the data projection can be considered the same as the linear multi-view embedding with respect to $\mathcal{H}_v$, and we only need an additional optimization solved by the SGD for updating the parameters of the networks. An additional constraint to have uni-variant projection can also be imposed as

$$\sum_{i=1}^{V} \mathbf{W}_i^\top \mathcal{H}_i \, \mathbf{L} \, \mathcal{H}_i^\top \mathbf{W}_i = \mathbf{I}. \tag{3.20}$$

We use the above constraint in Deep Multi-view CCA (DMvCCA). The objective of Deep Multi-view PLS (DMvPLS) is optimized with a constraint to have unit variance $\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{W}_i = \mathbf{I}$, while in Deep Multi-view Modular Discriminant Analysis (DMvMDA), we project the within-class scatter o the identity, i.e. we apply a form of whitening

$$\sum_{i=1}^{V} \mathbf{W}_i^\top \mathcal{H}_i \mathbf{L}_W \mathcal{H}_i^\top \mathbf{W}_i = \mathbf{I} \tag{3.21}$$

Using the variance constraint, the gradients in DMvCCA and DMvPLS become

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathcal{H}_i} &= \frac{\partial}{\partial \mathcal{H}_i} \operatorname{Tr} \left(\sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i^\top \mathcal{H}_i \, \mathbf{L} \, \mathcal{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathcal{H}_j \, \mathbf{L}, \end{aligned} \tag{3.22}$$

and the gradient of DMvMDA is expressed by

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathcal{H}_i} &= \frac{\partial}{\partial \mathcal{H}_i} \operatorname{Tr} \left(\sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i^\top \mathcal{H}_i \, \mathbf{L}_B^* \, \mathcal{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathcal{H}_j \, \mathbf{L}_B^*, \end{aligned} \tag{3.23}$$

Detailed derivation of (3.22) and (3.23) can be found in [P1].

3.1.5 Results and Discussion

3.1.5.1 Results on Zero-shot Recogntion

We present recognition accuracy using different methods in Table 3.2. A description of the datasets can be found in the experiment section of [P1]. The linear projection results are shown in the first block, the second block presents the kernel-based methods, the results of neural network based methods are shown in the third block, and the last block

Table 3.2: RECOGNITION ACCURACY (%) on the AwA DATASET [P1].

Method	2 views	3 views	4 views
Proposed LMvCCA	55.86	75.88	82.01
Proposed LMvPLS	**58.52**	73.59	77.09
Proposed LMvMDA	55.85	**77.64**	**82.88**
Proposed SLMvDA	54.58	69.02	70.56
Proposed KapMvCCA	56.41	73.40	74.76
Proposed KapMvPLS	55.58	**74.40**	75.05
Proposed KapMvMDA	**57.19**	71.64	**75.63**
Proposed DMvCCA	51.25	71.12	82.27
Proposed DMvPLS	43.28	68.81	74.63
Proposed DMvMDA	**53.87**	**75.61**	**83.66**
MvDA [35]	49.95	68.55	70.00
GMA [86]	52.12	73.49	78.46
MULDA [87]	55.46	74.13	74.88
TMV-HLP [7]	-	73.50	80.50
DCCA2 [32]	50.47	-	-

provides comparative results from methods in the literature. LMvCCA perform favorably comparing to other linear methods while the leading recognition rates can be found in the nonlinear methods using neural nets with 4 views. We adopt the explicit kernel mappings using random projections due to the large number of samples. However, the results are inferior compared to linear methods due to the information loss in sampling [88].

In terms of the detailed performance, the 4-view DMvMDA is reported to have the best result for zero-shot recognition. We also observe all methods consistently obtain a better accuracy with more views. Specifically in linear methods, LMvPLS has the highest accuracy with two input views. while LMvMDA provides a more discriminantive representation in the latent space leading to a better recognition when more views are presented. Nonlinear methods using neural networks were inferior to linear method in 2 and 3 views, but the model fitting improves significantly with enriched data from more views.

3.1.5.2 Results on Cross-modal Image Retrieval

Multi-modal and cross-modal image retrieval are enabled by using the proposed multi-view embedding method shown in Figure 3.3. Two image-to-text pairs are chosen randomly as queries, to perform image-to-image retrieval using both the *VGG-16* visual feature and the projected visual feature by the 4-view DCCA. We also perform text-to-image retrieval by querying the corresponding captions of the query image used in CBIR in the last column. We observe the CBIR performance can be improved by embedding the semantic information. Cross-modal image retrieval also provides a satisfying precision. More results comparing related methods can be found in Table V and VI in [P1].

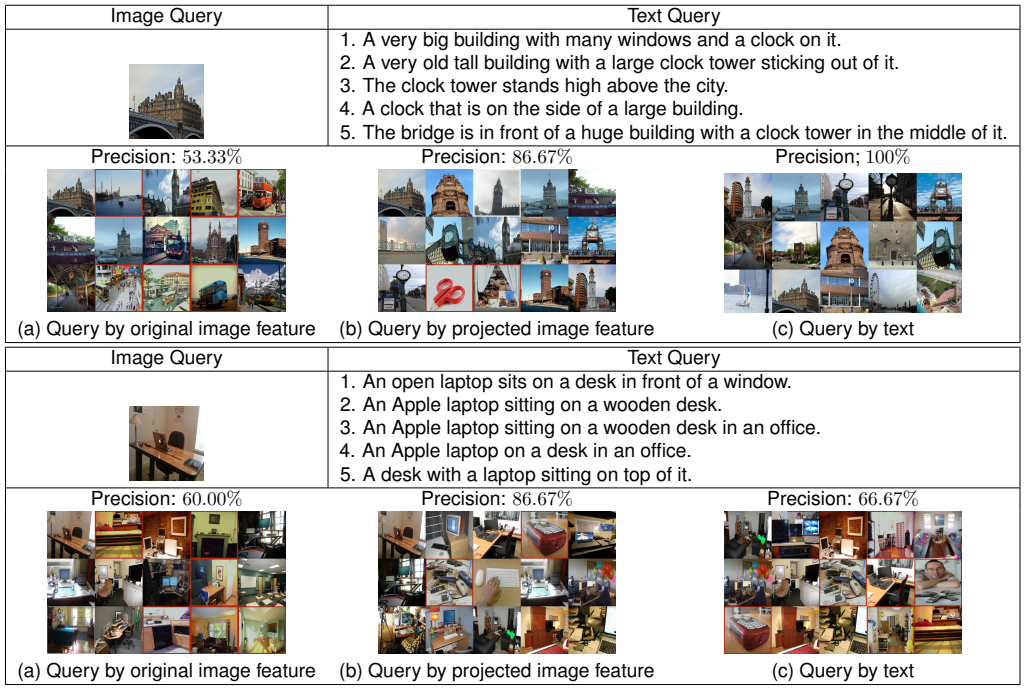


Figure 3.3: Sample retrieval results on the COCO dataset. The first row of each table presents the query image and text, and the second row shows the retrieved images by different query types. False positive results are bounded in red [P1].

3.2 Multi-view Nonparametric Discriminant Analysis

In the previous section, we introduced a group of multi-view embedding methods under the assumption that classes follow unimodal Gaussian distributions. Here a new criterion for multi-view discriminant analysis is formulated to enable larger number of projection directions by relaxing this assumption. It follows the same graph embedding framework as before; the between-class and within-class scatters are modeled by two different k-nearest neighbor graphs. We make use of all the samples in the intrinsic and penalty graphs, and class discrimination is encoded in sub-classes of neighboring samples which overpasses the assumption about the Gaussian distribution. A weighting scheme of neighboring sample pairs based on their proximity to the class boundary is introduced, which improves the feature discriminability in the latent space.

The criterion for multi-view nonparametric discriminant analysis (MvNDA) is

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^N)}{\text{Tr}(\mathbf{S}_W^N)}, \tag{3.24}$$

where $\mathbf{S}_B^N, \mathbf{S}_W^N$ denote the nonparametric between-class and within-class scatter matrices, respectively. The data projection $\mathbf{W}$ can be obtained by solving the generalized eigenvalue problem in (3.2). The intrinsic graph is designed to strengthen the intra-view

class compactness from the subclasses by using the k_1-nearest neighbors, and the within-class scatter matrix is

$$\mathbf{S}_W^N = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i (\mathbf{D}_W - \mathbf{V}_W) \mathbf{X}_i^\top \mathbf{W}_i \tag{3.25}$$

where $\mathbf{L}_W^N = \mathbf{D}_W - \mathbf{V}_W$ is the within-class Laplacian matrix and the intrinsic graph $\mathbf{V}_W$ is defined as

$$\mathbf{V}_{pq}^W = \begin{cases} 1, & \text{if } p \in \mathsf{NN}_{k_1}(q) \text{ or } q \in \mathsf{NN}_{k_1}(p) \\ 0, & \text{otherwise.} \end{cases} \tag{3.26}$$

$\mathsf{NN}_{k_1}(p)$ denotes the index set of the k_1 nearest neighbors of the sample $\mathbf{x}_p$ in the same class.

The view-specific penalty graph is formulated to encode the variance of the marginal samples from different classes of the same view as follows:

$$\mathbf{S}_B^{\mathsf{VS}} = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i [\mathbf{Q} \circ (\mathbf{D}_B - \mathbf{V}_B)] \mathbf{X}_i^\top \mathbf{W}_i, \tag{3.27}$$

where $\mathbf{L}_B^{\mathsf{VS}} = \mathbf{D}_B - \mathbf{V}_B$ is the between-class view-specific Laplacian matrix, and its intrinsic graph is characterized as:

$$\mathbf{V}_{pq}^B = \begin{cases} 1, & \text{if } (p,q) \in \mathsf{NP}_{k_2}(c_p) \text{ or } (p,q) \in \mathsf{NP}_{k_2}(c_q) \\ 0, & \text{otherwise.} \end{cases} \tag{3.28}$$

$\mathsf{NP}_{k_2}(c)$ is a set of data pairs which contains the k_2 nearest pairs in the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$. The weight matrix $\mathbf{Q}$ enhances the feature discrinability by strenghthening the importance of the samples on the classification boundary. Specifically, the value in $\mathbf{Q}$ goes to 0.5 if the sample falls close to the boundary, but reduces to 0 otherwise. $d(p,q)$ is the Euclidean distance between two vectors p and q. $\mathbf{Q}$ is given by:

$$\mathbf{Q}_{pq} = \begin{cases} \frac{\min\{d(p,q), d(p, \mathsf{NN}_{k_2}(p))\}}{d(p,q) + d(p, \mathsf{NN}_{k_2}(p))} & \text{if } (p,q) \in \mathsf{NP}_{k_2}(c_p) \\ & \text{or } (p,q) \in \mathsf{NP}_{k_2}(c_q) \\ 0 & \text{otherwise.} \end{cases}$$

The penalty graph is a linear combination of $\mathbf{S}_B^P$ of MvDA (2.30) and $\mathbf{S}_B^{\mathsf{VS}}$ of (3.27) to enforce both inter-view and intra-view class discrimination

$$\mathbf{S}_B^N = \alpha\, \mathbf{S}_B^P + (1-\alpha)\, \mathbf{S}_B^{\mathsf{VS}}, \tag{3.29}$$

where $\alpha \in [0,1]$ is a weighting factor which is set close to 1 if the training data has a Gaussian distribution, and some other value if the data distribution is unknown.

The intrinsic and penalty graphs are illustrated qualitatively in Figure 3.4. The within-class compactness is enforced by connecting a sample to its k_1-nearest-neighbors of the same

class and view. The between-class separability is also enforced by both connecting marginal point pairs from the same view but of different classes, and by using the distance of different class centers.

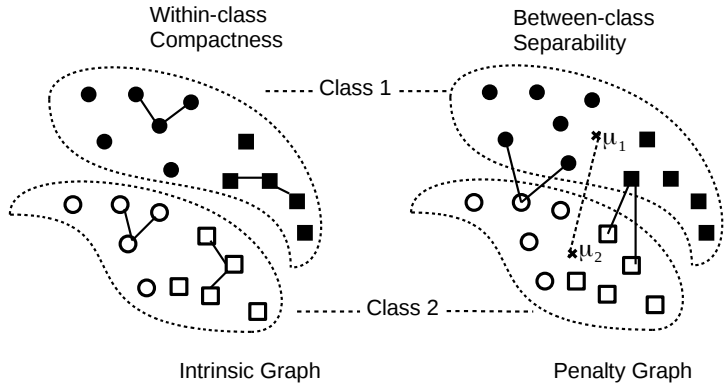


Figure 3.4: The adjacency relationship of the intrinsic and penalty graphs of the proposed MvNDA. The circular and rectangle dots indicate samples from different views. We illustrate the 2-nearest adjacencies (i.e. $k_1 = k_2 = 2$) of one sample in each class per view origin for clarity [P2].

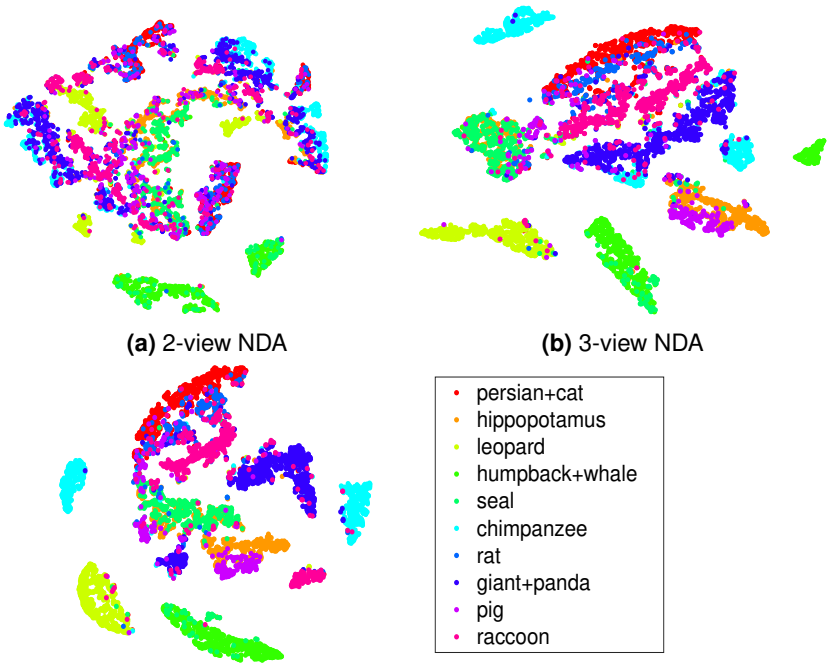
We also extend to nonlinear multi-view projections by employing kernel-based mapping. The input feature is mapped to the kernel space $\mathcal{F}_v$ using a nonlinear function ϕ, i.e. $\mathbf{X}_v \in \mathbb{R}^{D_v \times N} \xrightarrow{\phi(\cdot)} \mathbf{\Phi}(\mathbf{X}_v) \in \mathbb{R}^{|\mathcal{F}_v| \times N}$. In $\mathcal{F}_v$, according to the Representer Theorem [18, 19], the data projection matrix can be expressed as $\mathbf{W}_v = \mathbf{\Phi}(\mathbf{X}_v)\mathbf{A}_v$ and dot products between data pairs are represented using the kernel matrix $\mathbf{K}_v = \mathbf{\Phi}(\mathbf{X}_v)^\top \mathbf{\Phi}(\mathbf{X}_v)$ [17]. Then,

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{A}) = \arg\max_{\mathbf{A}} \frac{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}_B^N\, \mathbf{K}\mathbf{A})}{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}_W^N\, \mathbf{K}\mathbf{A}\,)}, \tag{3.30}$$

where $\mathbf{L}_B^N = \alpha \mathbf{L}_B^P + (1-\alpha)\mathbf{L}_B^{\text{VS}}$ is the between-class Laplacian matrix, and $\mathbf{K} = \text{diag}(\mathbf{K}_1, \cdots, \mathbf{K}_V)$. If the direct solution of (3.30) is impractical with a large size of the training data, we use the approximate kernel mapping proposed in [89] followed by the linear mapping defined in (3.24).

3.2.1 Results and Discussion

We visualize the embedded feature by the nonparametric linear multi-view discriminant analysis on AwA dataset in Figure 3.5. It can be seen that by increasing the number of views, the latent vector progresses from being distributed incoherently to showing more distinct groups associated with their corresponding semantic classes. Further comparisons on Wikipedia dataset between the various multi-view embedding methods are shown in Figure 3.6. The 4-view kernel-based methods are presented, and we observe that MvNDA is among the leading group in cross-modal image retrieval. Moreover, numerical analysis on the effects of different numbers of nearest neighbors and the weight factors α is illustrated in Figure 3.6. It shows the retrieval performs consistently



(a) 2-view NDA **(b)** 3-view NDA

(c) 4-view NDA

Figure 3.5: t-SNE Embedding of Latent Feature Representation: We visualize the embeddings from different numbers of views using the proposed method [P2].

with the different values of k or α, while only using the view-specific discrimination ($\alpha = 0$) degrades the MAP score.

3.3 Dropout Regularization for Linear Multi-view Subspace Learning

Dropout was originally proposed for regularizing the weights of neural network models by training an ensemble of sub-networks with certain neurons removed from an baseline network. Inspired by this technique, we propose a novel method to regularize the linear multi-view embedding. We define the data vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. We also define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}, v = 1, \ldots, V$ for the feature vectors of the vth view. The representation of each view samples in the latent space is given by a linear projection

$$\mathbf{Y}_v = \mathbf{W}_v^\top \mathbf{X}_v. \tag{3.31}$$

Inspired by dropout regularization, we apply an iterative optimization process. We create a binary mask $\mathbf{m}_{i,t}$ to remove the effects of certain values from the original feature vector in each epoch t. The elements in $\mathbf{m}_{i,t}$ equals to 1 with a probability value p following a Bernoulli distribution, and equals to 0 with $(1-p)$ probability. The resulting feature vector

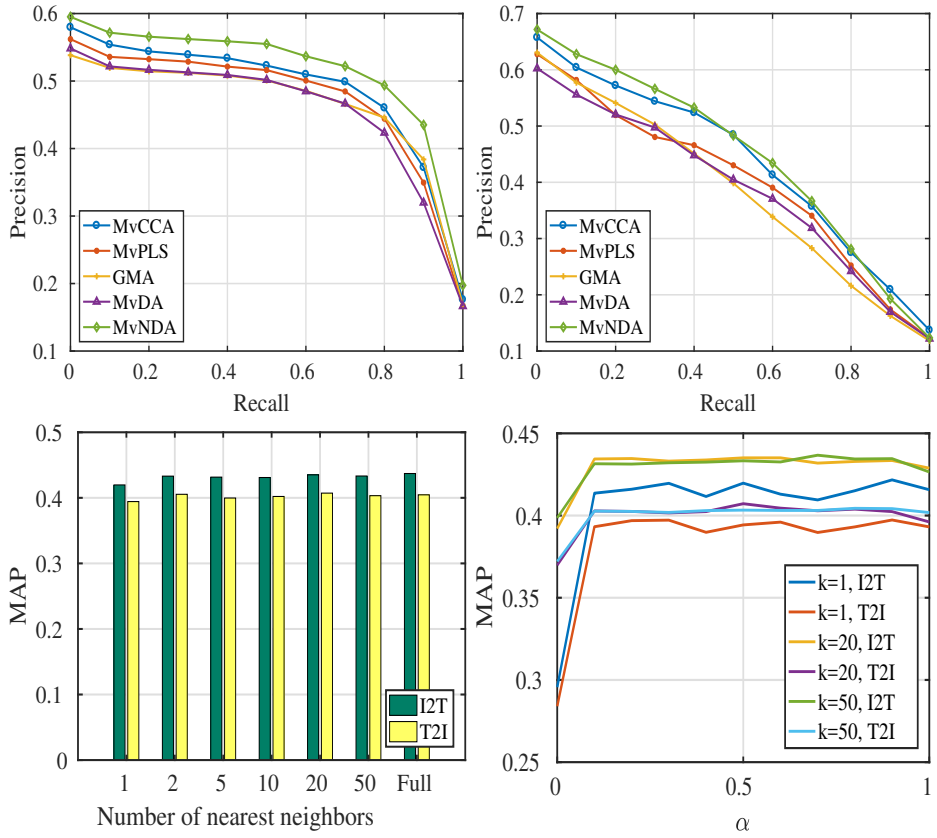


Figure 3.6: Clockwise from top left: The precision-recall curve by querying images for text annotations, the retrieval performance of matching text to images, the MAP scores with various α under different fixed numbers of nearest neighors k, (here $k = k_1 = k_2$), and the MAP scores with the different k nearest neighbors and a fixed $\alpha = 0.5$. The legends in the figures in the first row indicate the method producing the PR curve, and we denote querying images for texts by "I2T", and querying texts by images by "T2I" in the figure in the bottom row. k is the number of nearest neighbors [P2].

after applying dropout becomes

$$\mathbf{x}_{i,t} = \mathbf{m}_{i,t} \circ \mathbf{x}_i, \tag{3.32}$$

where $\circ$ denotes the operation for the Hadamard (element-wise) product. In order to achieve a robust mapping, we aim to have the representations of the original samples $\mathbf{x}_i$ and the masked versions of them $\mathbf{x}_{i,t}$ in the latent space as close as possible. By using the notation $\widetilde{\mathbf{x}}_{i,t} = \mathbf{x}_i - \mathbf{x}_{i,t}$ the above objective can be expressed as follows

$$\mathbf{W}^\top(\mathbf{x}_{v,i} - \mathbf{x}_{v,i,t}) = \mathbf{W}^\top \widetilde{\mathbf{x}}_{v,i,t} = \mathbf{0}. \tag{3.33}$$

Expressing the above for all training samples across the various views and iterations, we obtain the regularization term

$$R(\mathbf{W}) = \frac{1}{2\,N_T}\sum_{v=1}^{V}\sum_{i=1}^{N}\sum_{t=1}^{N_T}\|\mathbf{W}_v^\top\mathbf{x}_{v,i} - \mathbf{W}_v^\top\mathbf{x}_{v,i,t}\|_F^2 \tag{3.34}$$

$$= \frac{1}{2\,N_T}\sum_{v=1}^{V}\sum_{t=1}^{N_T}\|\mathbf{W}_v^\top\widetilde{\mathbf{X}}_{v,t}\|_F^2 \tag{3.35}$$

When the number of epoches N_T goes to infinity, based on the weak law of large numbers, we know that $R(\mathbf{W})$ will converge to its expected value

$$R(\mathbf{W}) = \frac{1}{2\,N_T}\sum_{v=1}^{V}\sum_{t=1}^{N_T}E\left(\mathbf{W}_v^\top\widetilde{\mathbf{X}}_{v,t}\widetilde{\mathbf{X}}_{v,t}^\top\mathbf{W}_v\right) \tag{3.36}$$

$$= \frac{1}{2\,N_T}\sum_{v=1}^{V}\sum_{t=1}^{N_T}\mathbf{W}_v^\top\left(\widetilde{\mathbf{X}}_{v,t}\widetilde{\mathbf{X}}_{v,t}^\top\circ\mathbf{P}\right)\mathbf{W}_v, \tag{3.37}$$

where $\mathbf{P} = \Big[(\mathbf{p}\mathbf{p}^\top)\circ(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\Big] + \Big[(\mathbf{p}\mathbf{I}^\top)\circ\mathbf{I}\Big]$, and $\mathbf{p} = [(1-p),\ldots,(1-p)]^\top \in \mathbb{R}^N$ is a vector whose elements shows the probability that $\mathbf{x}_i = 0$. We also define that $\mathbf{1} \in \mathbb{R}^{N\times N}$ as a vector of ones, and $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix.

Thus, the objective function for the dropout-based regularized linear multi-view subspace learning is

$$\begin{aligned}\mathcal{J} &= \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\frac{\mathrm{Tr}(\mathbf{S}_B)}{\mathrm{Tr}(\mathbf{S}_W + \alpha R(\mathbf{W}))}\\ &= \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\frac{\mathrm{Tr}(\mathbf{W}^\top\mathbf{P}\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top\mathbf{Q}\mathbf{W} + \alpha R(\mathbf{W}))}.\end{aligned} \tag{3.38}$$

where $\mathbf{S}_B$ is the between-class scatter matrix defined in (3.11) and $\mathbf{S}_W$ is the within-class scatter matrix of (3.13). $\mathbf{P}$ and $\mathbf{Q}$ are the inter-view and intra-view covariance matrices. α is the parameter adjusting the importance of regularization. The objective function integrates the inter-view and intra-view similarities to the dropout regularization. It is a modified form of Rayleigh quotient and can be solved as the generalized eigenvalue problem in (3.2).

3.3.1 Results on Face-Sketch Recognition

We demonstrate the effectiveness of using dropout regularization in cross-modal recognition in this section. Table 3.3 shows the accuracy for both face-sketch and sketch-face recognition by linear multi-view embedding methods including MvCCA, MvPLS, MvDA [35], and the proposed LDA without and with the dropout regularization. We observe that the MvMDA together with dropout regularization in the last row outperforms the relative methods by a large margin. It shows the feature discriminative power and robustness of

Table 3.3: Recognition Rate (%) on the CUFSF Dataset [P3].

Method	Face-Sketch	Sketch-Face	Avg.
MvCCA	48.79	52.83	50.81
MvPLS	31.38	31.38	31.38
MvDA	45.55	49.60	47.58
MvMDA	47.17	51.62	49.40
MvMDA-Dropout	61.13	64.98	**63.06**

the new method against over-fitting. Moreover, we also study the influence of the probability p on the recognition performance at different levels of regularization importance (α) in Figure 3.7. It can be seen that the recognition rate is generally consistent to different dropout probabilities, and always better than the one without the regularization, i.e. $p = 1$.

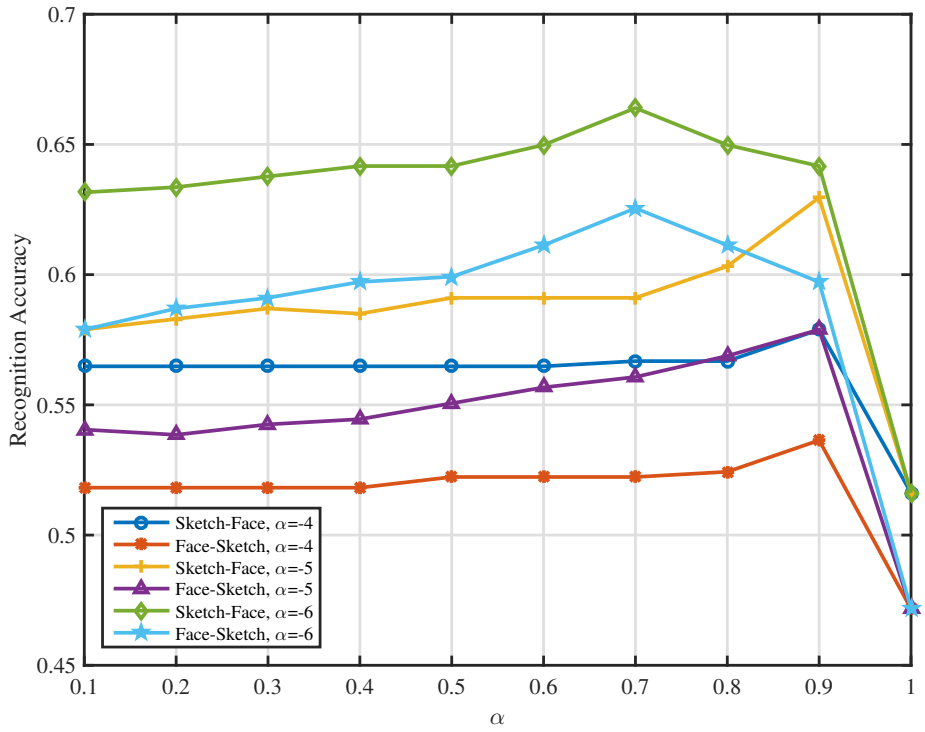


Figure 3.7: Face-Sketch Recognition Rate for different probability p [P3].

3.4 Multi-view Learning to Ranking

Learning to rank from multiple data sources is a relatively new problem in the area of data mining. We develop a novel way of ranking multi-facet objects. A typical example of such is found in university ranking. Figure 3.8 shows that several attributes in the THE

dataset [90], including teaching, research, student staff ratio and student number are highly correlated with all of the attributes in the ARWU dataset [91]. It enables finding a composite ranking by exploiting the correlation between individual rankings. We can observe that the indicators from different agencies may partially overlap and have a high correlation between each other.

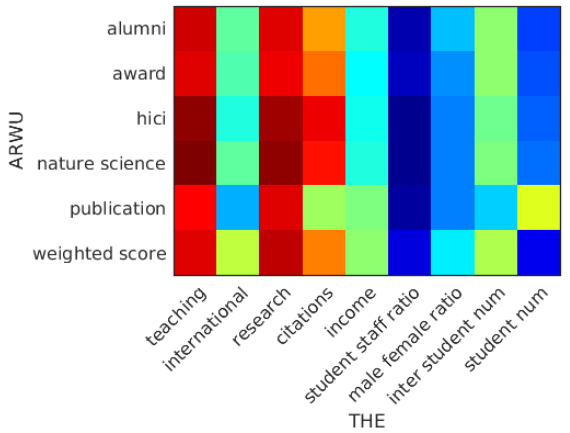


Figure 3.8: The correlation matrix between the measurements of Times Higher Education (THE) and Academic Ranking of World Universities (ARWU) rankings. The data is extracted and aligned based on the performance of the common universities in 2015 between the two ranking agencies. The reddish color indicates high correlation, while the matrix elements with low correlation are represented in bluish colors [P4].

We describe a novel composite ranking method which also keeps a close correlation with the individual rankings simultaneously. A multi-objective solution to ranking is introduced by capturing the information of the feature mapping from both within each view as well as across views using autoencoder-like networks. Moreover, we present a novel end-to-end solution to enhance the joint ranking with minimum view-specific ranking loss, so that the maximum global view agreements within a single optimization process is achieved. In the following sections, we will firstly describe the multi-view subspace learning to rank (MvSL2R), and then its specific formulations of MvCCAE and MvMDAE are presented. Finally, the end-to-end ranking solution is described.

3.4.1 Multi-view Subspace Learning to Rank (MvSL2R)

One straightforward way of multi-view learning to rank is to use the feature embeddings for ranking. The projected features in the common subspace are adopted to train a scoring function. The training data is generated from the intersection of ranking samples between views to have the same samples but various representations from different view origins. The overall ranking agreement is made by calculating the average voting from

the intersected ranking orders as

$$\overline{\mathbf{r}} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{r}_v. \tag{3.39}$$

The training input $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ of V views is obtained by performing the pairwise transform on the raw data, and the cross-view relevance scores $\overline{\mathbf{y}}$ is the average ranking orders $\overline{\mathbf{r}}$. We predict the relevance of new sample pairs using the probability function of the scoring function

$$\mathbf{p}_v(\mathbf{X}_v) = \frac{1}{1+\exp(-\mathbf{a}^\top \mathbf{W}_v^\top \mathcal{F}_v(\mathbf{X}_v))}, \tag{3.40}$$

where $\mathbf{W}_v$ is the data projection matrix of the vth view, and $\mathbf{a}$ is the weight from the logistic regressor described in (2.35). We summarize these steps in the algorithm below.

Algorithm 1: Multi-view Subspace Learning to Rank [P4].

1 Function MvSL2R $(\mathbf{X}, \mathbf{Y}, k)$;
  Input : The feature vectors of V views $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$, the relevance $\mathbf{y}$, and the dimensionality in the subspace k.
  Output : The predicted relevance probabilities $p = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_V\}$ of the new data.
2 Train a neural network to update the low-dimensinonal representation representation $\mathbf{Z}_v$ e.g. in (3.45) and (3.50). and the projection matrix $\mathbf{W} = [\mathbf{W}_1\mathbf{W}_2\ldots\mathbf{W}_V]^\top$.
3 Train a logistic regressor (2.35) as the scoring function to obtain the weight matrix $\mathbf{a}$.
4 Predict the new sample pairs for their relevance probabilities using (3.40) with the trained sub-networks $\mathcal{F}$ and $\mathcal{G}$, and the obtained weights $\mathbf{W}$ and $\mathbf{a}$.

3.4.2 Multi-view Canonically Correlated Auto-Encoder (MvCCAE)

A multi-objective solution to multi-view ranking is proposed by maximizing the between-view correlation while minimizing the reconstruction error from each view source, which is largely different from DMvCCA and DMvMDA, where only the nonlinear correlation between multiple views is optimized. Given the data matrix $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ of V views, the encoding network $\mathcal{F}$ and the decoding network $\mathcal{G}$, and the projection matrix $\mathbf{W}$, the objective of MvCCAE is formulated as follows,

$$\mathcal{J}_{\text{MvCCAE}} = \arg\max\ \mathcal{J}'_{\text{DMvCCA}} - \alpha\sum_{v}^{V}\ell_{\text{AE}}\Big(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))\Big), \tag{3.41}$$

where we introduce the new objective

$$\mathcal{J}'_{\mathsf{DMvCCA}} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max} \frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1\\ j\neq i}}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_j^\top\mathbf{W}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_i^\top\mathbf{W}_i\right)}, \tag{3.42}$$

and the loss function of the vth autoencoder is $\ell_{\mathsf{AE}}(\mathbf{X}_v;\mathcal{G}_v(\mathcal{F}_v(\cdot))) = \|\mathbf{X}_v - \mathcal{G}_v(\mathcal{F}_v(\mathbf{X}_v))\|_2 + \rho\sum_l \|\nabla_{\mathbf{X}_v}\mathcal{F}_v^l(\mathbf{X}_v)\|_2$, with the L^2 regularization at the lth intermediate layer of the vth view denoted by $\mathbf{Z}_v^l = \mathcal{F}_v^l(\mathbf{X}_v)$. Here, α and ρ are controlling parameters for the trade-off between the terms. In the ranking problems, we devise a new method to directly optimize the Rayleigh quotient criterion in (3.1) and let

$$f = \mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j\neq i\\ j=1}}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_j^\top\mathbf{W}_j\right),$$

and

$$g = \mathrm{Tr}\left(\sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_i^\top\mathbf{W}_i\right).$$

Here, the output of each sub-network $\mathcal{F}_v$ is denoted by $\mathbf{Z}_v = \mathcal{F}_v(\mathbf{X}_v)$. Then, we have

$$\frac{\partial f}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V}\sum_{\substack{j\neq i\\ j=1}}^{V}\mathbf{W}_i\,\mathbf{W}_j^\top\mathbf{Z}_j\,\mathbf{L}, \tag{3.43}$$

and

$$\frac{\partial g}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V}\mathbf{W}_i\,\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}. \tag{3.44}$$

By using (3.43) and (3.44) and following the quotient rule, we derive the stochastic optimization of MvCCAE to be

$$\begin{aligned}\frac{\partial \mathcal{J}_{\mathsf{MvCCAE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g\,\frac{\partial f}{\partial \mathbf{Z}_v} - f\,\frac{\partial g}{\partial \mathbf{Z}_v}\right)\\ - \frac{\partial}{\partial \mathbf{Z}_v}\alpha\sum_{v}^{V}\ell_{\mathsf{AE}}(\mathbf{X}_v;\mathcal{G}_v(\mathcal{F}_v(\cdot))).\end{aligned} \tag{3.45}$$

The gradient to compute the autoencoding loss ℓ_{AE} is derived from the view-specific sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$. The sub-network $\mathcal{F}_v$ is optimized with $\frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v}$ to obtain the output $\mathbf{Z}_v$, while the gradient of $\mathcal{G}_v$ network with respect to its parameters can be obtained using the chain rule from $\frac{\partial \mathcal{G}_v(\mathbf{X}_v)}{\partial \mathbf{Z}_v}$.

3.4.3 Multi-view Modularly Discriminant Auto-Encoder (MvMDAE)

Similar to MvCCAE, the objective of MvMDAE is to optimize the view-specific reconstruction error and the cross-view correlation as follows,

$$\mathcal{J}_{\text{MvMDAE}} = \arg\max \mathcal{J}'_{\text{DMvMDA}} - \alpha \sum_{v}^{V} \ell_{\text{AE}}\Big(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))\Big). \tag{3.46}$$

We define a new objective for the cross-view correlation

$$\mathcal{J}'_{\text{DMvMDA}} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max} \frac{\text{Tr}\big(\sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_B \mathbf{Z}_j^\top \mathbf{W}_j\big)}{\text{Tr}\big(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_W \mathbf{Z}_i^\top \mathbf{W}_i\big)}, \tag{3.47}$$

3.4.3.1 Optimization

The detailed optimization is derived by replacing the laplacian matrix in MvCCAE with $\mathbf{L}_B$ and $\mathbf{L}_W$. We let

$$f = \text{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j\neq i\\ j=1}}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_B \, \mathbf{Z}_j^\top \mathbf{W}_j\right),$$

and

$$g = \text{Tr}\left(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_W \, \mathbf{Z}_i^\top \mathbf{W}_i\right).$$

Then, we have

$$\frac{\partial f}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V}\sum_{\substack{j\neq i\\ j=1}}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathbf{Z}_j \, \mathbf{L}_B, \tag{3.48}$$

and

$$\frac{\partial g}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V} \mathbf{W}_i \, \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_W. \tag{3.49}$$

The stochastic optimization of MvMDAE can be derived by using (3.48), (3.49) and applying the quotient rule as follows,

$$\begin{aligned}\frac{\partial \mathcal{J}_{\text{MvMDAE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g \, \frac{\partial f}{\partial \mathbf{Z}_v} - f \, \frac{\partial g}{\partial \mathbf{Z}_v}\right) \\ - \frac{\partial}{\partial \mathbf{Z}_v} \alpha \sum_{v}^{V} \ell_{\text{AE}}(\mathbf{X}_v, \mathcal{G}_v(\mathcal{F}_v(\cdot))).\end{aligned} \tag{3.50}$$

The gradient of the objective can be calculated using the chain rule, and the stochastic gradient descent (SGD) is used with mini-batches for optimization.

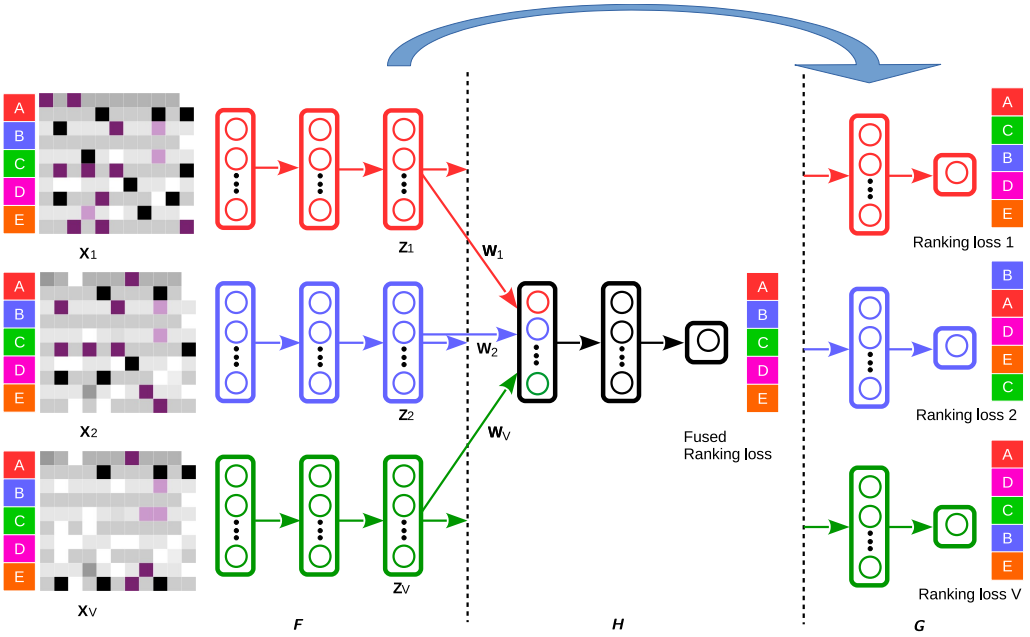


Figure 3.9: System diagram of the Deep Multi-view Discriminant Ranking (DMvDR). First, the features $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ are extracted for data representations in different views and fed through the individual sub-network $\mathcal{F}_v$ to obtain the nonlinear representation $\mathbf{Z}_v$ of the vth view. The results are then passed through two pipelines of networks. One line goes to the projection $\mathbf{W}$, which maps all $\mathbf{Z}_v$ to the common subspace, and their concatenation is trained to optimize the fused ranking loss with the fused sub-network $\mathcal{H}$. The other line connects $\mathbf{Z}_v$ to the sub-network $\mathcal{G}_v, \forall v = 1, \ldots, V$ for the optimization of the vth ranking loss [P4].

3.4.4 Deep Multi-view Discriminant Ranking (DMvDR)

Multi-view Subspace Learning to Rank devises a multi-objection solution, while it does not have a direct connection to ranking. Alternatively, we propose another end-to-end method to optimize the view-specific and the joint ranking together in the single network as shown in Figure 3.9. University ranking can be used as an example, where different lists are generated from different ranking agencies, and each agency has a different set of measurements. Given the inputs $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$, the cross entropy loss (2.36) is optimized with the view-specific relevance $\mathbf{y}$ during training and the joint view relevance $\overline{\mathbf{y}}$. The attributes $\mathbf{X}_v$, where $v = 1, \ldots, V$, are trained through the view-specific sub-network $\mathcal{F}_v$ based on their own evaluation metric. We exploit the intermediate representation from the neural networks $\mathbf{Z}_v = \mathcal{F}_v(\mathbf{X}_v), v = 1, \ldots, V$, which are the inputs of the joint network $\mathcal{H}$ as $\mathbf{W}_v^\top \mathbf{Z}_v$, $v = 1, \ldots, V$, after the mappings to generate the joint university ranking list. Each of nonlinear embeddings is also the input to the view-specific network $\mathcal{G}_v$, which minimizes its distance to the original ranking $\mathbf{r}_v$. In contrast to MvCCAE and MvMDAE in learning an embedding from the bottleneck of the neural network, we similarly exploit the effectiveness of intermediate layers $\mathbf{Z}_v$ in between the view-specific sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$, but towards the ranking loss for DMvDR. The detailed procedure of this method

is described below.

The gradient of each view-specific sub-network $\mathcal{G}_v$ is calculated from the output $\mathbf{y}$ with respect to its parameters. Since the loss passes from each view-specific $\mathcal{F}_v$ to $\mathcal{G}_v$ sub-network, the gradient can be calculated independently with respect to the output of each view-specific $\mathcal{F}_v$ sub-network as $\frac{\partial \mathbf{y}}{\partial \mathbf{Z}} = \{\frac{\partial \mathbf{y}_1}{\partial \mathbf{Z}_1}, \frac{\partial \mathbf{y}_v}{\partial \mathbf{Z}_v}, \ldots, \frac{\partial \mathbf{y}_V}{\partial \mathbf{Z}_V}\}$ Then, the gradient of $\frac{\partial \mathbf{y}_v}{\partial \mathcal{G}_v}$ with respect to its network weights can be determined through backpropagation [92]. All sub-networks contain several layers with Sigmoid functions.

The fused sub-network $\mathcal{H}$ is updated with the gradient of the ranking loss from the cross-view relevance scores $\overline{\mathbf{y}}$. Similar to the generation of training data in MvSL2R, we find the intersection of the ranking data with different representations or measurements from various sources, and perform the pairwise transform to have the sample pairs as the input $\mathcal{X}$ and $\overline{\mathbf{y}}$ from the cross-view ranking orders $\overline{\mathbf{r}}$ in (3.39). As a result, the input $\mathbf{S}$ to the fused sub-network $\mathcal{H}$ is the concatenation of the nonlinear mapping from the V view-specific networks $\mathcal{F}_v$ as

$$\mathbf{S} = [\mathbf{W}_1^\top \mathbf{Z}_1 \;\; \mathbf{W}_2^\top \mathbf{Z}_2 \;\; \ldots \;\; \mathbf{W}_V^\top \mathbf{Z}_V]^\top. \tag{3.51}$$

We develop two possible scenarios during testing: (a) For the aligned testing samples, the results from nonlinear mappings are concatenated in the same manner as the training phase to generate a fused ranking list $\overline{\mathbf{p}}$ at the end of common $\mathcal{H}$ sub-network; and (b) if we have missing samples or completely missing views in testing, then we use $\mathbf{S} = \mathbf{W}_v^\top \mathbf{Z}_v$ for the vth view. Note that the resulting view-specific prediction $\mathbf{p}_v$ still maintains the cross-view agreement which is ranked from the trained joint network. The gradient of $\frac{\partial \overline{\mathbf{y}}}{\partial \mathbf{S}}$ and $\frac{\partial \overline{\mathbf{y}}}{\partial \mathcal{H}}$ can be easily calculated afterwards using the SGD.

A multi-view subspace embedding layer is developed and further trained for joint ranking. The input to the sub-network $\mathcal{H}$ is a concatenation of the projected features of the outputs from the sub-networks $\mathcal{F}_v$. The gradient of multi-view subspace embedding (MvSE) can be derived from (3.48) and (3.49):

$$\frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g\,\frac{\partial f}{\partial \mathbf{Z}_v} - f\,\frac{\partial g}{\partial \mathbf{Z}_v}\right). \tag{3.52}$$

We forward pass the gradient from the embedding layer to the fused sub-network $\mathcal{H}$. The embedding layer acts like a hub of the network as the layers of $\mathcal{F}_v$ is backward optimized through it. Furthermore, it also influences the parameters in $\mathcal{G}_v$ for their view-specific ranking loss.

The update of the common view-specific $\mathcal{F}_v$ depends both on the view-specific ranking output $\mathbf{y}$ and the cross-view relevance $\overline{\mathbf{y}}$. The v-th sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$ are optimized consecutively using backpropagation with respect to the gradient $\frac{\partial \mathbf{y}}{\partial \mathbf{X}_v}$. At the same time, the loss with respect to the fused ranking $\overline{\mathbf{y}}$ is passed through multi-view subspace

embedding (MvSE) from $\mathbf{S}$ in (3.51) as the input to the fused sub-network $\mathcal{H}$. The resulting gradient of each sub-network $\mathcal{F}_v$ is given by

$$\begin{aligned}\frac{\partial \mathcal{J}_{\text{DMvDR}}}{\partial \mathbf{Z}_v} =& \frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} - \alpha \sum_{v}^{V} \frac{\partial}{\partial \mathbf{Z}_v} \ell_{\text{Rank}}(\mathbf{X}_v, \mathbf{y}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))) \\ & - \beta \frac{\partial}{\partial \mathbf{Z}_v} \ell_{\text{Rank}}(\mathbf{S}, \overline{\mathbf{y}}; \mathcal{H}(\cdot)),\end{aligned} \tag{3.53}$$

where α and β are the scaling factors controlling the magnitude of the ranking loss. The gradients with respect to their parameters can be obtained by following the chain rule similar to the other sub-networks.

We summarize the update of the entire network of DMvDR using the SGD with mini-batches below, and denote the parameters of the sub-network as $\theta = \{\theta_{\mathcal{F}_1}, \theta_{\mathcal{F}_2}, \ldots, \theta_{\mathcal{F}_V}, \theta_{\mathcal{G}_1}, \theta_{\mathcal{G}_2}, \ldots, \theta_{\mathcal{G}_V}, \theta_{\mathcal{H}}\}$. A gradient descent step is $\Delta\theta = -\eta \frac{\partial}{\partial \theta} \mathcal{J}_{\text{DMvDR}}$, where η is the learning rate. The gradient update step at time t can be summarize according to the chain rule as follows

$$\begin{aligned}\Delta\theta^t =& \{\Delta\theta^t_{\mathcal{F}_1}, \Delta\theta^t_{\mathcal{F}_2}, \ldots, \Delta\theta^t_{\mathcal{F}_V}, \\ & \Delta\theta^t_{\mathcal{G}_1}, \Delta\theta^t_{\mathcal{G}_2}, \ldots, \Delta\theta^t_{\mathcal{G}_V}, \Delta\theta^t_{\mathcal{H}}\} \\ \Delta\theta^t_{\mathcal{G}_v} =& -\frac{\partial \ell_{\text{rank}}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathcal{G}_v} \\ \Delta\theta^t_{\mathcal{H}} =& -\frac{\partial \ell_{\text{rank}}}{\partial \overline{\mathbf{y}}} \cdot \frac{\partial \overline{\mathbf{y}}}{\partial \mathcal{H}} \\ \Delta\theta^t_{\mathcal{F}_v} =& \frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v} - \frac{\partial \ell_{\text{rank}}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v} \\ & - \frac{\partial \ell_{\text{rank}}}{\partial \overline{\mathbf{y}}} \cdot \frac{\partial \overline{\mathbf{y}}}{\partial \mathbf{S}} \cdot \frac{\partial \mathbf{S}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v}.\end{aligned} \tag{3.54}$$

Data is generated by the pairwise transform for both training and testing. The test samples are evaluated based on the relative relevance to their queries. Meanwhile, it is also possible to feed the raw ranking data into the trained model to predict their overall ranking positions.

3.4.5 Experiments and Discussion on University Ranking

We present the ranking performance of the proposed multi-view learning to rank methods on university ranking problems. Related methods in subspace learning and co-training methods are included as follows for comparison. The subspace learning methods follow the MvSL2R method proposed in Section 3.4.1 for ranking.

- **Best Single View**: a method which shows the best performance of Ranking SVM [78] over the individual views.

- **Feature Concat**: a method which concatenate the features of the common samples for training a Ranking SVM [78].
- **LMvCCA [5]**: a linear multi-view CCA method.
- **LMvMDA [5]**: a linear supervised method for multi-view subspace learning.
- **MvDA [35]**: another linear supervised method for multi-view subspace learning. It differs from the above in that the view difference is not encoded in this method.
- **SmVR [41]**: a semi-supervised method that seeks a global agreement in ranking. It belongs to the category of co-training. We develop the complete data in the following experiments for training so that its comparison with the subspace learning methods is fair. Therefore, SmVR becomes a supervised method in this paper.
- **DMvCCA [5]**: a nonlinear extension of LMvCCA using neural networks.
- **DMvMDA [5]**: a nonlinear extension of LMvMDA using neural networks.
- **MvCCAE**: the first proposed multi-view subspace learning to rank method proposed in the paper.
- **MvMDAE**: the supervised multi-view subspace learning to rank method proposed in the paper.
- **DMvDR**: the end-to-end multi-view learning to rank method proposed in the paper.

We show a rank correlation matrix of plots in Figure 3.10 with correlations among pairs of ranking lists from the views 1-3 and the predicted list denoted by 'Fused'. Histograms of the ranking data are shown along the matrix diagonal, while scatter plots of data pairs appear off diagonal. We calculated the slopes of the least-squares reference lines in the scatter plots from the displayed correlation coefficients. We generated the fused ranking list by the proposed DMvDR from the common universities in 2015. From the correlations between the views 1-3, we observe that the correlation coefficients are generally low, with the highest (0.81) between view 1 and 3, while the others are around 0.70. In contrast, the fused rank has a high correlation to each view. The scatter plots and the reference lines are well aligned, and the correlation coefficients are all above 0.80, demonstrating that the proposed DMvDR effectively exploits the global agreement with all view.

Moreover, we also report the average prediction results over 3 different university datasets of the proposed and competing methods in Table 3.4. Due to the misalignment of ranking data in 2015 across datasets, we make the ranking prediction based on the individual view input, which is described in details in the Section 3.4.4. Ranking SVM [78] on the single feature or its concatenation performs poorly compared to the other methods. It shows that when the data is heterogeneous, the feature concatenation cannot enhance

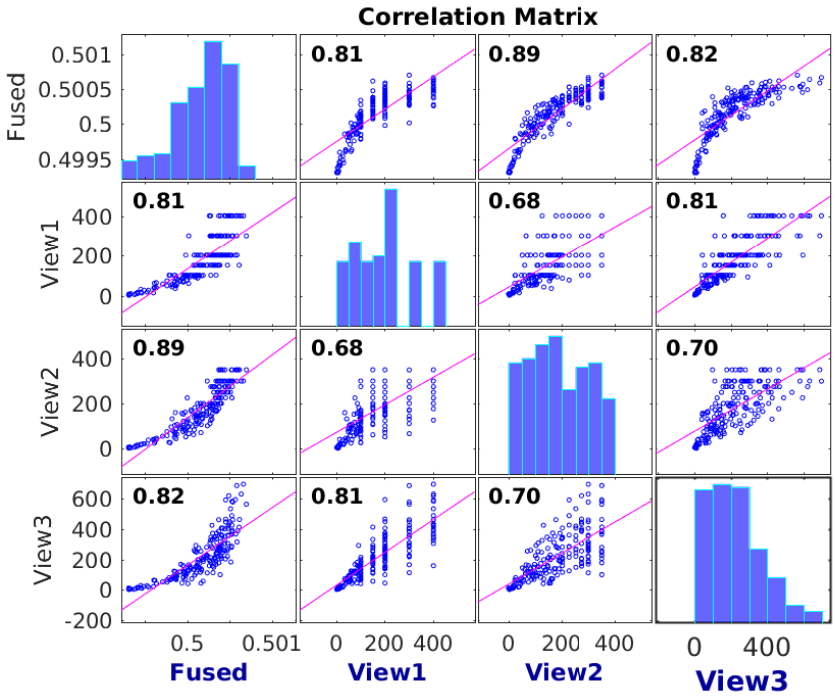


Figure 3.10: Rank correlation matrix for views 1-3 and the fused view [P4].

Table 3.4: Average Prediction Results (%) on 3 University Ranking Datasets in 2015 [P4].

Methods	Kendal's tau	Accuracy
Best Single View	65.38	-
Feature Concat	35.10	-
LMvCCA [5]	86.04	94.49
LMvMDA [5]	87.00	94.97
MvDA [35]	85.81	94.34
SmVR [41]	80.75	-
DMvCCA [5]	70.07	93.20
DMvMDA [5]	70.81	94.75
MvCCAE (*ours*)	75.94	94.01
MvMDAE (*ours*)	81.04	94.85
DMvDR (*ours*)	**89.28**	**95.30**

joint ranking. Kendal's tau from the linear subspace learning methods are comparatively higher than their nonlinear counterparts. This is due to the fact that the nonlinear methods aim to maximize to the correlation in the embedding space, while the scoring function is not optimized for ranking. Finally, DMvDR has the highest ranking and classification performance using its end-to-end optimization process.

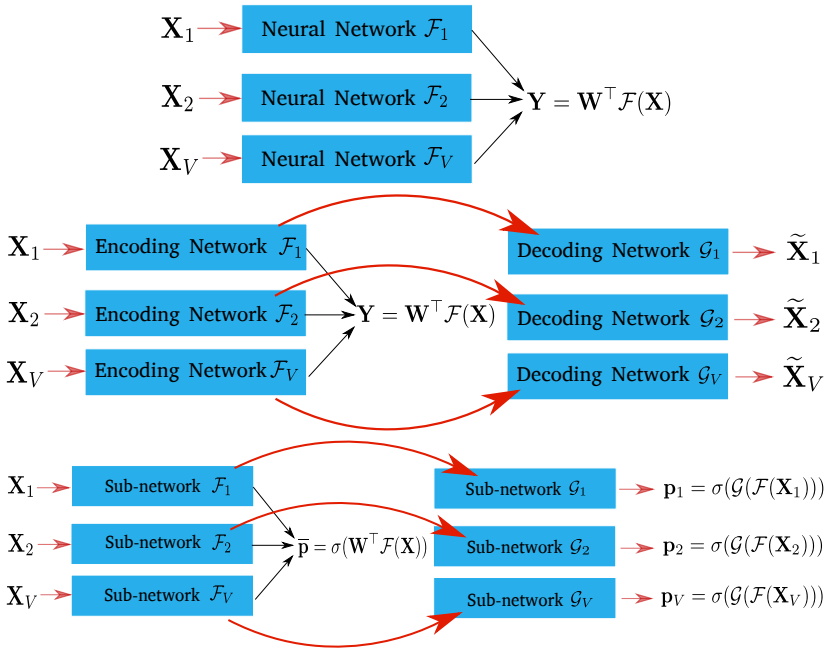


Figure 3.11: A summary of multi-view deep learning methods.

3.5 Contribution to Multi-view Deep Learning

We can summarize our contribution to multi-view deep neural networks in three groups of methods shown in Figure 3.11. The first group merges the feature embeddings at the end of each neural network. The second class of feature learning methods exploits autoencoders and concatenates the features in between (sub-)networks. Both view-specific reconstruction loss and the joint loss are optimized. The last category is the end-to-end solution for direct decision making (ranking specifically in the thesis). The compositive objective and the view-specific losses are optimized together towards to the objective rankings.

4 Conclusion

Multi-view data analysis is an important and active research topic in the field of visual recognition and data mining. We are entering a world of heterogeneous data, which urges extensive studies on unifying data representations from multiple sources to enhance feature discriminability. Semantic gap has been mitigated by exploiting the multi-view embeddings in various domains and feature types. Moreover, cross-modal matchings are enabled using the methods presented in the thesis.

Multi-view learning was reviewed in Chapter 2. Subspace learning finds a common latent space from different sensory modalities by fitting an optimization criterion. The compact and discriminant feature is used for cross-modal multimedia retrieval, zero-shot object recognition, face-sketch recognition and learning to rank. Regularization methods were reviewed in particularly dropout and dropconnect to alleviate the outfitting problem in neural networks. Multi-view learning methods in the literature were described extensively including multi-view subspace learning and co-training. Unsupervised learning techniques including Canonical Correlation Analysis, Partial Least Squares regression. Their nonlinear correspondences were also briefly reviewed. Multi-view discriminant analysis and several of its extensions were included in the thesis as supervised methods.

The methods developed throughout the thesis were presented in Chapter 3. The generalized multi-view embedding method using the graph embedding framework was introduced. Multi-view CCA, PLS and LDA were shown to be characterized by their specific intrinsic and penalty graph matrices within the same framework. Multi-view Modular Discriminant Analysis was proposed by exploiting the distances between class centers of different views. Meanwhile, nonlinear embeddings were studied, together with implicit and explicit kernel mappings for multi-view learning. A unified scheme for learning by neural networks was developed which combined the learned representations with a linear embedding layer. The stochastic gradient descent for optimizing the proposed objective function were formulated.

Secondly, a multi-view nonparametric discriminant analysis method were formulated by using two different KNN graphs to encode view discrepancy and weighting the contribution of neighboring pairs based on their proximity to the class boundary.The novel method

allows for multiple projecion directions, by relaxing the Gaussian distribution assumpton of related methods. Additionally, a dropout regularization for linear multi-view subspace learning was introduced and demonstrated its effectiveness in overcoming the overfitting problem in cross-modal recognition. Finally, novel deep multi-view learning to rank methods were presented which can provide a composite ranking method while keeping a close correlation with the individual rankings simultaneously. The proposed methods were multi-objective solutions to ranking by capturing the information of the feature mapping from within each view as well as across views. Moreover, intermediate representations were exploited using either autoencoders or discriminant learning. The end-to-end solution presented in the thesis is able to enhance the joint ranking with minimum view-specific ranking loss within a single optimization process.

To conclude, the thesis addressed several challenging multi-view data analysis problems by learning representations using the unified solution for multi-view embedding. Proposed methods have shown promising performance in object recognition, cross-modal image retrieval, face recognition and object ranking. In the future, we should further explore the reduction of computational complexity for multi-view kernel methods for big data. Learning from incomplete and unlabeled multi-view data should be studied for video analysis. Potential applications in automated driving and video surveillance systems can be explored to broaden the scope of applications.

References

[1] X. Wang, X. Wang, and D. Wilkes, "A divide-and-conquer approach for minimum spanning tree-based clustering," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 21, no. 7, pp. 945 –958, July 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. [Online]. Available: http://mscoco.org/home/

[3] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.

[4] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[5] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2542–2555, Sept 2018.

[6] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with cnn visual features: A new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, 2017.

[7] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.

[8] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville, "Guesswhat?! visual object discovery through multi-modal dialogue," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, vol. abs/1611.08481, 2017.

[9] X. Wang and X. Tang, “Face photo-sketch synthesis and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, Nov 2009.

[10] G. Cao, M. A. Waris, A. Iosifidis, and M. Gabbouj, “Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval,” in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[11] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, “Vqa: Visual question answering,” *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[12] I. Jolliffe, *Principal Component Analysis*. Springer, New York, NY, 1986.

[13] X. Wang and X. Tang, “A unified framework for subspace face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 26, no. 9, pp. 1222–1228, 2004.

[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[15] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: a general framework for dimensionality reduction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[16] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*., Oct 2007, pp. 1–7.

[17] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, “An introduction to kernel-based learning algorithms,” *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[18] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, “Input space versus feature space in kernel-based methods,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in *Proceedings of Annual Conference of Computational Learning Theory,*. Springer, Heidelberg, Germany, 2001, pp. 416–426.

[20] A. Sharma and D. W. Jacobs, “Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 593–600.

[21] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 23, no. 2, pp. 228–233, 2001.

[22] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[23] A. Iosifidis, A. Tefas, and I. Pitas, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

[24] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 27, no. 3, pp. 328–340, 2005.

[25] T. Hastie, R. Tibshirani, and J. Friedmani, *The Elements of Statistical Learning*. Springer, 2013, ch. 3 Linear Methods for Regression, pp. 61–79.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2018.

[29] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[30] M. Borga, "Canonical correlation: a tutorial," http://people.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf, 2001.

[31] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[32] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1247–1255.

[33] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4590–4594.

[34] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic press, 1980, ch. 10 Canonical Correlation Analysis, pp. 281–290.

[35] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[36] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern recognition*, vol. 34, no. 7, pp. 1405–1416, 2001.

[37] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[38] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*. ACM, 2000, pp. 86–93.

[39] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in neural information processing systems*, 2011, pp. 1413–1421.

[40] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 1713–1720.

[41] N. Usunier, M.-R. Amini, and C. Goutte, "Multiview semi-supervised learning for ranking multilingual documents," *Machine Learning and Knowledge Discovery in Databases*, pp. 443–458, 2011.

[42] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 22, no. 12, pp. 1349 –1380, 12 2000.

[43] F. Alaya Cheikh, "Muvis: A system for content-based image retrieval," *F. Alaya Cheikh, PhD. Thesis at Tampere University of Technology, Tampere, Finland*, 2004.

[44] S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Computer Vision and Image Understanding*, vol. 75, no. 1-2, pp. 86–98, 1999.

[45] M. J. Swain, C. Frankel, and V. Athitsos, "Webseer: An image search engine for the world wide web," in *International Conference on Computer Vision and Pattern Recognition*, 1997.

[46] C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia tools and applications*, vol. 25, no. 1, pp. 5–35, 2005.

[47] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, 2008.

[48] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 2006, pp. 321–330.

[49] T. Tsikrika and J. Kludas, "Overview of the wikipedia multimedia task at imageclef 2009," in *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2009, pp. 60–71.

[50] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 175–184.

[51] Y.-T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, 2008.

[52] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, 2008.

[53] Y. Fu, T. Xiang, Y. G. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 112–125, Jan 2018.

[54] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.

[55] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.

[56] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Multi-task deep neural network for joint face recognition and facial attribute prediction," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 365–374.

[57] V. Ferrari and A. Zisserman, "Learning visual attributes," in *Advances in Neural Information Processing Systems*, 2008, pp. 433–440.

[58] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 503–510.

[59] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.

[60] D. Parikh and K. Grauman, “Interactively building a discriminative vocabulary of nameable attributes,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 1681–1688.

[61] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, “Learning multimodal latent attributes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 303–316, 2014.

[62] X. Tang and X. Wang, “Face sketch recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50–57, Jan 2004.

[63] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, “A nonlinear approach for face sketch synthesis and recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, vol. 1. IEEE, 2005, pp. 1005–1010.

[64] B. F. Klare and A. K. Jain, “Heterogeneous face recognition using kernel prototype similarities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 35, no. 6, pp. 1410–1422, 2013.

[65] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, “Coupled discriminant analysis for heterogeneous face recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1707–1716, 2012.

[66] Y. Jin, J. Lu, and Q. Ruan, “Coupled discriminative feature learning for heterogeneous face recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, 2015.

[67] C. Peng, X. Gao, N. Wang, and J. Li, “Graphical representation for heterogeneous face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[68] J. Yu, D. Tao, M. Wang, and Y. Rui, “Learning to rank using user clicks and visual features for image retrieval,” *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[69] X. Li, T. Pi, Z. Zhang, X. Zhao, M. Wang, X. Li, and P. S. Yu, “Learning bregman distance functions for structural learning to rank,” *IEEE Transactions on Knowledge and Data EngineeringS*, vol. 29, no. 9, pp. 1916–1927, Sept 2017.

[70] O. Wu, Q. You, X. Mao, F. Xia, F. Yuan, and W. Hu, “Listwise learning to rank by exploring structure of objects,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1934–1939, 2016.

[71] Y. Zhu, G. Wang, J. Yang, D. Wang, J. Yan, J. Hu, and Z. Chen, "Optimizing search engine revenue in sponsored search," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 588–595.

[72] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Picture collage," *IEEE Transactions on Multimedia (TMM)*, vol. 11, no. 7, pp. 1225 –1239, 2009.

[73] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[74] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems*, 1998, pp. 451–457.

[75] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 89–96.

[76] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," 2000.

[77] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[78] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[79] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[80] J. Xu and H. Li, "AdaRank: a boosting algorithm for information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 391–398.

[81] C. J. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions," in *Advances in neural information processing systems*, 2007, pp. 193–200.

[82] F. Feng, L. Nie, X. Wang, R. Hong, and T.-S. Chua, "Computational social indicators: A case study of chinese university ranking," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2017, pp. 455–464.

[83] H.-J. Ye, D.-C. Zhan, Y. Miao, Y. Jiang, and Z.-H. Zhou, "Rank consistency based multi-view learning: a privacy-preserving approach," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 991–1000.

[84] W. Gao and P. Yang, "Democracy is good for ranking: Towards multi-view rank learning and adaptation in web search," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 63–72.

[85] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[86] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[87] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, Dec 2016.

[88] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[89] A. Iosifidis and M. Gabbouj, "Nyström-based approximate kernel subspace learning," *Pattern Recognition*, vol. 57, pp. 190–197, 2016.

[90] "The times higher education world university ranking," https://www.timeshighereducation.com/world-university-rankings, 2016.

[91] N. C. Liu and Y. Cheng, "The academic ranking of world universities," *Higher education in Europe*, vol. 30, no. 2, pp. 127–136, 2005.

[92] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 1998, pp. 9–50.

Publications

Publication I

G. Cao, A. Iosifidis, K. Chen and M. Gabbouj, “Generalized Multi-View Embedding for Visual Recognition and Cross-Modal Retrieval,” in IEEE Transactions on Cybernetics, vol. 48, no. 9, pp. 2542-2555. doi: 10.1109/TCYB.2017.2742705

Generalized Multi-view Embedding for Visual Recognition and Cross-modal Retrieval

Guanqun Cao, Alexandros Iosifidis, *Senior Member, IEEE*, Ke Chen and Moncef Gabbouj, *Fellow, IEEE*
{guanqun.cao, ke.chen, moncef.gabbouj}@tut.fi, alexandros.iosifidis@eng.au.dk

***Abstract*—In this paper, the problem of multi-view embedding from different visual cues and modalities is considered. We propose a unified solution for subspace learning methods using the Rayleigh quotient, which is extensible for multiple views, supervised learning, and non-linear embeddings. Numerous methods including Canonical Correlation Analysis, Partial Least Square regression and Linear Discriminant Analysis are studied using specific intrinsic and penalty graphs within the same framework. Non-linear extensions based on kernels and (deep) neural networks are derived, achieving better performance than the linear ones. Moreover, a novel Multi-view Modular Discriminant Analysis (MvMDA) is proposed by taking the view difference into consideration. We demonstrate the effectiveness of the proposed multi-view embedding methods on visual object recognition and cross-modal image retrieval, and obtain superior results in both applications compared to related methods.**

I. Introduction

People see the world differently, and objects are described from various point of views and modalities. Identifying an object can not only benefit from visual cues including color, texture and shape, but textual annotations from different observations and languages. Thanks to data enrichment from sensor technologies, the accuracy in image retrieval and recognition has been significantly improved by taking advantage of multi-view and cross-domain learning [1], [2]. Since matching the data samples across various feature spaces directly is infeasible, subspace learning approaches, which learn a common feature space from multi-view spaces, becomes an effective approach in solving the problem.

Numerous methods have been proposed in subspace learning. They can be grouped into three major categories based on the characteristics of machine learning: *two-view learning* and *multi-view learning*; *unsupervised learning* and *supervised learning*; and *linear learning* and *non-linear learning*. While traditional techniques in multivariate analysis take two inputs [3], multi-view methods have been proposed to find an optimal representation from more than two views [4], [5]. Compared to learning the feature transformation in an unsupervised manner, discriminative methods, such as Linear Discriminant Analysis (LDA) have been extended to multi-view cases. Additionally, the transformation can also be kernel-based or learned by (deep) neural nets to exploit their non-linear properties.

Two-view learning and *multi-view learning*: One of the most popular methods in multivariate statistics is Canonical Correlation Analysis (CCA) [6]. It seeks to maximize the correlation between two sets of variables. Alternatively, its multi-view counterpart aims to obtain a common space from $V > 2$ views [4], [5], [7]. This is achieved either by scaling the cross-covariance matrices to incorporate the covariances from more than two views, or by finding the best rank-1 approximation of the data covariance tensor. A similar approach to find the common subspace is Partial Least Square Regressions [8]. It maximizes the cross-covariance from two views by regressing the data samples to the common space. Besides transformation and regression, Multi-view Fisher Discriminant Analysis (MFDA) [9] learns the transformation minimizing the difference between data samples of predicted labels. The Dropout regularization was introduced for the multi-view linear discriminant analysis in [10].

Unsupervised learning and *supervised learning*: In contrast to unsupervised transformations, including CCA and PLS, LDA [11], [12] exploits the class labels effectively by maximizing the between-class scatter while minimizing the within-class scatter simultaneously. CCA has been successfully combined with LDA to find a discriminative subspace in [13], [14], [15]. Coupled Spectral Regression (CSR) [16] projects two different inputs to the low-dimensional embedding of labels by PLS regressions. Consistent with the original LDA, a Multi-view Discriminant Analysis (MvDA) [17] finds a discriminant representation over V views. The between-class scatter is maximized regardless of the difference between inter-view and intra-view covariances, while the within-class scatter is minimized in the mean time. Generalized Multi-view Analysis (GMA) [18] was proposed to maximize the intra-view discriminant information. Recently, a semi-supervised alternative [19] was also proposed for multi-view learning, which adopts a non-negative matrix factorization method for view mapping and a robust sparse regression model for clustering the labeled samples. Moreover, a multi-view information bottleneck method [20] was proposed to retain its discrimination and robustness for multi-view learning.

Linear and *non-linear learning*: Many problems are not linearly separable and thereby kernel-based methods and learning representation by (deep) neural nets are introduced. By mapping the features to the high dimensional feature space using the kernel trick [21], kernel CCA [22] adopts a pre-

The authors are with the Laboratory of Signal Processing, Tampere University of Technology, Finland. A. Iosifidis is also with the Dept. of Engineering, Electrical and Computer Engineering, Aarhus University, DK-8200, Aarhus N, Denmark., Denmark.

This work was supported by the NSF-TEKES Center for Visual and Decision Informatics (CVDI), sponsored by Tieto Oy Finland. A. Iosifidis and K. Chen were supported from the Academy of Finland Postdoctoral Research Fellowships (No. 295854 and 298700, respectively).

defined kernel and limits its application on small datasets. Many linear multi-view methods subsequently made their kernel extension [23], [15], [24]. Kernel approximation [5] was adopted later to work on big data. Deep CCA [25] was proposed using neural nets to learn adaptive non-linear representations from two views, and uses the weights in the last layers to find the maximum correlation. A similar idea has been exploited on LDA [26]. PCANet [27] was introduced to adopt a cascade of linear transformation, followed by binary hashing and block histograms.

We make several contributions in this paper: First, we propose a unified multi-view subspace learning method for CCA, PLS and LDA techniques using the graph embedding framework [11]. We design both intrinsic and penalty graphs to characterize the intra-view and inter-view information, respectively. The intra-view and inter-view covariance matrices are scaled up to incorporate more than two views for numerous techniques by exploiting their specific intrinsic and penalty graphs. In our proposed Multi-view Modular Discriminant Analysis (MvMDA), the two graphs also charaterize the within-class compactness and between-class separability. Based on the aforementioned characteristics of subspace learning algorithms, we propose a generalized objective function for multi-view subspace learning using Rayleigh quotient. This unified multi-view embedding approach can be solved as a generalized eigenvalue problem.

Second, we introduce a Multi-view Modular Discriminant Analysis (MvMDA) method by exploiting the distances between centers representing classes of different views. This is of particular interest since the resulting scatter encodes cross-view information, which empirically is shown to provide superior results. Third, we also extend the unified framework to the non-linear cases with kernels and (deep) neural networks. Kernel-based multi-view learning method is derived with an implicit kernel mapping. For larger datasets, we use the explicit kernel mapping [28] to approximate the kernel matrices. We also derive the formulation of stochastic gradient descent (SGD) for optimizing the objective function in the neural nets.

Last but not least, we demonstrate the effectiveness of the proposed embedding methods on visual object recognition and cross-modal image retrieval. Specifically, zero-shot recognition is evaluated by discovering novel object categories based on the underlying intermediate representation [29], [30], [31]. Its performance is heavily dependent on the representation in the latent space shared by visual and semantic cues. We integrate observations from *attributes* as a middle-level semantic property for the joint learning. Superior recognition results are achieved by exploiting the latent feature space with non-linear solutions learned from the multi-view representations. We also employ the proposed multi-view subspace learning methods for cross-modal image retrieval [1], [32], [**?**], [33]. This type of methods differs from the co-training methods for image classification [34] and web image reranking [35], [36]. In the experiments, we show promising retrieval results performed by embedding more modalities into the common feature space, and find that even conventional content-based image retrieval can be improved.

Fig. 1: Visualization of test images from the AwA dataset grouped by the features in the subspace. We highlight one of the representative classes "leopard" bounded in orange to show images of the same animal categories are positioned in their neighborhoods after multi-view embedding. Note the 2-dimensional t-SNE map [37] is generated from a near circular shape.

The rest of the paper is organized as follows. Section II reviews the related work. In Section III, we show the unified formulation to generalize the subspace learning methods. It is followed by the extension to multi-view techniques and derivation in kernels and neural nets. Then, in Section IV, we present the comparative results in zero-shot object recognition and cross-modal image retrieval on three popular multimedia datasets. Finally, Section V concludes the paper.

II. Related work

In this section, we first define the common notations used throughout the paper. Then, we will briefly review the related methods for multi-view subspace learning. Moreover, recent work on non-linear methods concerning kernels and (deep) neural networks are discussed.

A. Notations

We define the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. We also define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}, v = 1, \ldots, V$ for the feature vectors of the vth view, and discard the index in the single-view case for notation simplicity. Note that the dimensionality of the various feature spaces D_v may vary across the views. The covariance matrix is a statistics commonly used in CCA and PLS. We denote $\bar{\mathbf{X}}_v = \mathbf{X}_v - \frac{1}{N}\mathbf{X}_v \mathbf{e}\mathbf{e}^\top$ as the centered data matrix. The cross-view covariance matrix between view i and j is then expressed as $\mathbf{\Sigma}_{ij} = \frac{1}{N}\bar{\mathbf{X}}_i \bar{\mathbf{X}}_j^\top =$

$\frac{1}{N}\mathbf{X}_i\left(\mathbf{I}-\frac{1}{N}\mathbf{e}\,\mathbf{e}^\top\right)\mathbf{X}_j^\top$, where $\mathbf{e}\in\mathbb{R}^N$ is a vector of ones and $\mathbf{I}\in\mathbb{R}^{N\times N}$ is the identity matrix. For the supervised learning problems, the class label of the sample $\mathbf{x}_i$ is noted as $c_i\in\{1,2,\ldots,C\}$, where C is the number of classes. We define the class vector $\mathbf{e}^c\in\mathbb{R}^N$ with $e_c(i)=1$, if $c_i=c$, and $e_c(i)=0$, otherwise. $\mathbf{W}_v\in\mathbb{R}^{D_v\times d}, v=1,\ldots,V$ is the projection matrix for each view, d is the number of dimensions in the latent space. The feature dimension D_v in the original space of each view is usually high, which makes the distribution of the samples sparse, leading to several problems including the small sample size problem [38]. Therefore we want to project the samples to the latent space.

The generic projection function is defined to project $\mathbf{X}\in\mathbb{R}^{D\times N}$ to $\mathbf{Y}\in\mathbb{R}^{d\times N}$. We define the linear projection by $\mathbf{Y}=\mathbf{W}^\top\mathbf{X}$. In kernel methods, we map the data to a Hilbert space $\mathcal{F}$. Let us define $\phi(\cdot)$ as the non-linear function mapping $x_i\in\mathbb{R}^D$ to $\mathcal{F}$, and $\mathbf{\Phi}=[\boldsymbol{\phi}(\mathbf{x}_1),\ldots,\boldsymbol{\phi}(\mathbf{x}_N)]$ as the data matrix in $\mathcal{F}$. In multi-view cases, $\mathbf{\Phi}=[\mathbf{\Phi}_1^\top,\ldots,\mathbf{\Phi}_V^\top]^\top$. Since the dimensionality of $\mathcal{F}$ is arbitrary, the kernel trick [39] is exploited in order to implicitly map the data to $\mathcal{F}$. The Gram matrix is given by

$$\mathbf{K}_v=\kappa(\mathbf{X}_v,\mathbf{X}_v)=\mathbf{\Phi}_v^\top\cdot\mathbf{\Phi}_v, \tag{1}$$

where $\kappa(\cdot,\cdot)$ is the so-called kernel function. The centered Gram matrix is $\bar{\mathbf{K}}_v=\mathbf{K}_v-\frac{1}{N}\mathbf{1}\mathbf{K}_v-\frac{1}{N}\mathbf{K}_v\mathbf{1}^\top+\frac{1}{N^2}\mathbf{1}\mathbf{K}_v\mathbf{1}$, where $\mathbf{1}\in\mathbb{R}^{N\times N}$ is an all-ones matrix. In order to find the optimal projection, we can express $\mathbf{W}_v$ of each view as a linear combination of the training samples in the kernel space based on the Representer Theorem [21], [40]. This can be expressed by using a new weight matrix $\mathbf{A}_v$ as

$$\mathbf{W}_v=\mathbf{\Phi}_v\mathbf{A}_v. \tag{2}$$

In the case where a neural network with M layers is considered, $\boldsymbol{\beta}_j$ contains the weight parameters in the jth layer, $j=1,\ldots,M$. The weights $\mathbf{B}=[\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_M]$ are learned by applying stochastic gradient descent (SGD), and $h(\cdot\,;\mathbf{B})$ is a non-linear mapping function which maps $\mathbf{X}_v$ to the representation of the last hidden layer $\mathbf{H}_v$, i.e.

$$\mathbf{H}_v=h(\mathbf{X}_v;\mathbf{B}_v), \tag{3}$$

where $\mathbf{B}_v$ is the weight matrix trained by applying backpropagation in the vth network.

B. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) [6], [41] is a conventional statistical technique which finds the maximum correlation between two sets of data samples $\mathbf{X}_1\in\mathbb{R}^{D_1\times N}$ and $\mathbf{X}_2\in\mathbb{R}^{D_2\times N}$ using the linear combination $\mathbf{Y}_1=\mathbf{W}_1^\top\mathbf{X}_1$ and $\mathbf{Y}_2=\mathbf{W}_2^\top\mathbf{X}_2$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are determined by optimizing:

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\operatorname{corr}(\mathbf{W}_1^\top\mathbf{X}_1,\mathbf{W}_2^\top\mathbf{X}_2) \tag{4}$$

$$=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\frac{\mathbf{W}_1^\top\mathbf{\Sigma}_{12}\mathbf{W}_2}{\sqrt{\mathbf{W}_1^\top\mathbf{\Sigma}_{11}\mathbf{W}_1}\cdot\sqrt{\mathbf{W}_2^\top\mathbf{\Sigma}_{22}\mathbf{W}_2}}, \tag{5}$$

where

$$\mathbf{\Sigma}=\begin{bmatrix}\mathbf{\Sigma}_{11}&\mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21}&\mathbf{\Sigma}_{22}\end{bmatrix}=\frac{1}{N}\begin{bmatrix}\bar{\mathbf{X}}_1\bar{\mathbf{X}}_1^\top&\bar{\mathbf{X}}_1\bar{\mathbf{X}}_2^\top\\ \bar{\mathbf{X}}_2\bar{\mathbf{X}}_1^\top&\bar{\mathbf{X}}_2\bar{\mathbf{X}}_2^\top\end{bmatrix} \tag{6}$$

C. Kernel CCA

Kernel CCA finds the maximum correlation between two views after mapping them to the kernel space [22]. This is expressed by

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}\operatorname{corr}(\mathbf{W}_1^\top\mathbf{\Phi}_1,\mathbf{W}_2^\top\mathbf{\Phi}_2) \tag{7}$$

We use the kernel trick [39] and the Representer Theorem in (2), and derive the objective function for the kernel CCA as

$$\mathcal{J}=\arg\max_{\mathbf{A}_1,\mathbf{A}_2}\frac{\mathbf{A}_1^\top\mathbf{K}_1\mathbf{K}_2\,\mathbf{A}_2}{\sqrt{\mathbf{A}_1^\top\mathbf{K}_1\mathbf{K}_1\,\mathbf{A}_1}\cdot\sqrt{\mathbf{A}_2^\top\mathbf{K}_2\mathbf{K}_2\,\mathbf{A}_2}}. \tag{8}$$

D. Deep CCA

Deep CCA maximizes the correlation between a pair of views by learning non-linear representations from the input data through multiple stacked layers of neurons [25], [42]. A linear CCA layer is added on top of both networks, and the inputs to the CCA layer depend on the network outputs $\mathbf{H}_1$ and $\mathbf{H}_2$. Similar to the non-linear case in (8), a modified objective function $\min_{\mathbf{W}_1,\mathbf{W}_2}-\frac{1}{N}\operatorname{Tr}\left(\mathbf{W}_1{}^\top\mathbf{H}_1\,\mathbf{H}_2^\top\mathbf{W}_2\right)$ is optimized, where $\mathbf{W}_1,\mathbf{W}_2$ are the projection matrices in the CCA layer, and the correlated outputs are $\mathbf{Y}_1=\mathbf{W}_1^\top\mathbf{H}_1$ and $\mathbf{Y}_2=\mathbf{W}_2^\top\mathbf{H}_2$. A modified SGD method is developed with respect to the inputs $\mathbf{H}_1$ and $\mathbf{H}_2$ to the linear layer, which are also the outputs from the two networks. The objective function is expressed as $\operatorname{Tr}\left(\mathbf{W}_1^\top\mathbf{H}_1\,\mathbf{H}_2^\top\mathbf{W}_2\right)=\operatorname{Tr}(\mathbf{T}^\top\mathbf{T})^{\frac{1}{2}}$, which describes the correlation as the sum of the top d singular vectors of $\mathbf{T}=\mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$ whose definition can be found in [3].

E. Partial Least Squares (PLS) regression

Partial Least Squares (PLS) regression [8] is another dimensionality reduction technique derived from the linear combination of the input vectors $\mathbf{X}_1$ together with the target information which is considered as the second view $\mathbf{X}_2$. PLS maximizes the between-view covariance by solving

$$\mathcal{J}=\arg\max_{\mathbf{W}_1,\mathbf{W}_2}[\operatorname{Tr}(\mathbf{W}_1^\top\mathbf{X}_1\mathbf{X}_2^\top\mathbf{W}_2)], \tag{9}$$

$$\text{subject to } \mathbf{W}_1^\top\mathbf{W}_1=\mathbf{I},\mathbf{W}_2^\top\mathbf{W}_2=\mathbf{I}. \tag{10}$$

The non-linear extensions of PLS are obtained in the similar manner as the ones in CCA.

F. Generalized Multi-view Analysis (GMA)

GMA [18] is a generalized framework incorporating numerous dimensionality reduction methods. It maximizes the intra-view discriminant information, but ignores the inter-view information.

$$\mathcal{J} = \arg\max_{\mathbf{W}} \left[\mathrm{Tr} \left(\sum_{i}^{V} \sum_{i<j}^{V} 2\lambda_{ij} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{X}_j^\top \mathbf{W}_j + \sum_{i=1}^{V} \mu_i \mathbf{W}_i^\top \mathbf{P}_i \mathbf{W}_i \right) \right], \quad \text{subject to } \sum_{i}^{V} \mathbf{W}_i^\top \mathbf{Q}_i \mathbf{W}_i = \mathbf{I}. \tag{11}$$

Here both $\mathbf{P}$ and $\mathbf{Q}$ are the intra-view covariance matrices. $\mathbf{P}$ is a square matrix and $\mathbf{Q}$ is a square symmetric definite matrix. We adopt Generalized Multiview Marginal Fisher Analysis (GMMFA) in this framework. The method is also kernelizable using the Representer Theorem and kernel trick.

G. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [11], [43] finds the projection by maximizing the ratio of the between-class scatter to the within-class scatter. Let us define by $\boldsymbol{\mu}_c$ the mean vector of the c'th class, formed by N_c samples, and $\boldsymbol{\mu}$ the global mean. Then, LDA optimizes the following criterion:

$$\mathcal{J} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{P}\, \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{Q}\, \mathbf{W})}, \tag{12}$$

where

$$\mathbf{P} = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top = \mathbf{X} \Big(\sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top - \frac{1}{N} \mathbf{e}\, \mathbf{e}^\top \Big) \mathbf{X}^\top, \tag{13}$$

$$\mathbf{Q} = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top = \mathbf{X} \Big(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top \Big) \mathbf{X}^\top. \tag{14}$$

Non-linear extensions with kernels include KDA [44] and KRDA [45].

H. Multi-view Discriminant Analysis (MvDA)

MvDA [17] is the multi-view verison of LDA which maximizes the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. Its objective function is

$$\mathcal{J} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{S}_B^M)}{\mathrm{Tr}(\mathbf{S}_W^M)}, \tag{15}$$

where the between-class scatter matrix is

$$\mathbf{S}_B^M = \sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \Big(\sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top - \frac{1}{N} \mathbf{e}\, \mathbf{e}^\top \Big) \mathbf{X}_j^\top \mathbf{W}_j, \tag{16}$$

and the within-class scatter matrix is

$$\mathbf{S}_W^M = \sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \Big(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top \Big) \mathbf{X}_j^\top \mathbf{W}_j. \tag{17}$$

$\mathbf{W}$ contains the eigenvectors of the matrix $\mathbf{S} = {\mathbf{S}_W^M}^{-1} \mathbf{S}_B^M$ corresponding to the leading d eigenvalues λ_i.

III. Generalized Multi-view Embedding

Here we propose a generalized expression of objective function for multi-view subspace learning. The generalized optimization problem is given by:

$$\mathcal{J} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{P} \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{Q} \mathbf{W})} \tag{18}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the matrices describing the inter-view and intra-view covariances, respectively. The above equation has the form of the Rayleigh quotient. Therefore, all subspace learning methods that maximize the criterion can be reduced to a generalized eigenvalue problem:

$$\mathbf{P}\mathbf{W} = \boldsymbol{\rho}\, \mathbf{Q}\mathbf{W}, \tag{19}$$

and the solution is given in (20) below:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_V \end{pmatrix} \text{ and } \boldsymbol{\rho} = \sum_{i=1}^{d} \lambda_i \tag{20}$$

are the generalized eigenvector and the sum of the top d generalized eigenvalues λ_i, respectively. $\mathbf{W}$ contains the projection matrices of all views, and $\boldsymbol{\rho}$ is the value of Rayleigh quotient. We address the Rayleigh quotient as the uniform objective function, reaching out to all subspace learning methods in the paper. The non-linear multi-view embeddings can be achieved by kernel mappings, or (deep) neural networks optimized by SGD. Suppose we have a linear projection $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$, $\mathbf{S}_{vij}$ is a similarity weight matrix which encodes the intra-view properties to be minimized, and $\mathbf{S}'_{vij}$ is a penalty weight expressing the inter-view properties to be maximized. Then based on [11], [46], we can express the objective function as follows

$$\mathcal{J} = \arg\max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \frac{\sum_{v=0}^{V} \sum_{i=0}^{N} \sum_{j=0}^{N} \mathbf{S}'_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2}{\sum_{v=0}^{V} \sum_{i=0}^{N} \sum_{j=0}^{N} \mathbf{S}_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2} \tag{21}$$

$$= \arg\max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L}' \mathbf{X}^\top \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L} \mathbf{X}^\top \mathbf{W})}. \tag{22}$$

In the kernel case, we also have

$$\mathcal{J} = \arg\max_{\mathbf{A}^\top \mathbf{K} \mathbf{A} = \mathbf{I}} \frac{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}' \mathbf{K} \mathbf{A})}{\mathrm{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}\, \mathbf{K} \mathbf{A})}. \tag{23}$$

In the above, we define the diagonal matrix of each view pair as $\mathbf{D}_{uv}$ whose i-th element is $[\mathbf{D}_{uv}]_{ii} = \sum_j [\mathbf{S}_{uv}]_{ij}$, and the total graph Laplacian matrix as $\mathbf{L} = \mathbf{D} - \mathbf{S}$. Similarly, we have $\mathbf{D}', \mathbf{S}', \mathbf{L}'$ in the penalty graph.

For the non-linear mapping by neural networks, we deploy a linear embedding layer on top of the networks. This scheme is illustrated in Fig. 2. Since we have more than two input views, we train multiple neural networks whose outputs are connected to the linear layer and the objective is the same as in the linear case. By backpropagating the error of the weight matrix, we optimize the Rayleigh quotient criterion with respect to the non-linear feature representation from each view in the last hidden layer of the networks. The projection is found in the same way as in the linear case, and we will address the SGD formulation for the specific algortihms in the next section.

Fig. 3 illustrates the proposed framework graphically. We can extract different types of low-level features from images, texts, and intermediate representations. The multi-modal feature vectors are passed through linear or non-linear projec-

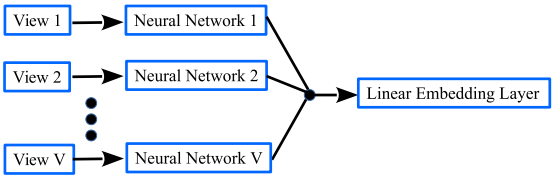


Fig. 2: An illustration of Multi-view (Deep) Embedding Neural Networks.

tions to the latent space. The projected features characterize the properties of the intra-view compactness and inter-view separability based on the proposed criterion. We show the scaled inter-view and intra-view matrices for each multi-view algorithm in the next section. Then, the projection matrices are presented with respect to their own intrinsic and penalty graph matrices and the optimization methods.

A. Scaling up the inter-view and intra-view covariance matrices

The idea behind multi-view CCA (MvCCA) is to maximize the correlation between all pairs of views. Its objective can be rephrased as maximizing the inter-view covariance while minimizing the intra-view covariance in the latent space. Therefore, we consider inter-view covariance matrices between different view representations in $\mathbf{P}$ and the covariance matrices of each view in $\mathbf{Q}$. Multi-view PLS (MvPLS) maximizes the inter-view covariance directly. Since we also embed the target information for the subspace learning, the proposed MvPLS differs from MvCCA only in the intra-view minimization. Taking the class discrimination into consideration, the novel multi-view modular discriminant analysis (MvMDA) extends to separate the data of different classes between views while making the intra-class data compact. We illustrate the structure of $\mathbf{P}$ and $\mathbf{Q}$ for each method in Table I.

TABLE I: The matrices $\mathbf{P}$ and $\mathbf{Q}$ for the proposed multi-view CCA, PLS and MvMDA.

	$\mathbf{P}$	$\mathbf{Q}$
MvCCA	$\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1V} \\ \mathbf{\Sigma}_{21} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{V1} & \mathbf{\Sigma}_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{VV} \end{bmatrix}$
MvPLS	$\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1V} \\ \mathbf{\Sigma}_{21} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{V1} & \mathbf{\Sigma}_{V2} & \cdots & \mathbf{0} \end{bmatrix}$	$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$
MvMDA	$\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1V} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{V1} & \mathbf{P}_{V2} & \cdots & \mathbf{P}_{VV} \end{bmatrix}$	$\begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{VV} \end{bmatrix}$

B. Linear subspace learning

When the subspace projection is linear, we can obtain the latent feature vectors from each view as

$$\mathbf{Y}_v = \mathbf{W}_v^\top \mathbf{X}_v, \tag{24}$$

and the projection matrix is derived directly by solving the generalized eigenvalue problem in (19). As shown in Table I, multi-view CCA has the total covariance matrix $\mathbf{\Sigma} = \mathbf{P} + \mathbf{Q}$, and we derive its projection matrix by fulfilling the criterion below

$$\mathcal{J} = \underset{\mathbf{W}_v, v=1,\ldots,V}{\arg\max} \frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_i^\top \mathbf{W}_i\right)}, \tag{25}$$

where the Laplacian matrix $\mathbf{L} = \mathbf{I} - \frac{1}{N}\mathbf{e}\,\mathbf{e}^\top$.

Multi-view PLS has the same Laplacian matrix as the one in Multi-view CCA. We only optimize the Rayleigh quotient by maximizing the cross-covariance matrices between different views as

$$\mathcal{J} = \underset{\mathbf{W}^\top \mathbf{W} = \mathbf{I}}{\arg\max}\, \mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1 \\ j\neq i}}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L} \mathbf{X}_j^\top \mathbf{W}_j\right), \tag{26}$$

whose solution is the projection matrix.

We propose two ways to determine the projection matrix in multi-view LDA. The first appoach is the multi-view extension of the standard LDA, and its between-class scatter $\mathbf{S}_B$ maximizes the distance between the class means from all views:

$$\begin{aligned} \mathbf{S}_B &= \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{p=1}^{C}\sum_{\substack{q=1 \\ p\neq q}}^{C} (\boldsymbol{m}_p^i - \boldsymbol{m}_q^j)(\boldsymbol{m}_p^i - \boldsymbol{m}_q^j)^\top \\ &= \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_B \mathbf{X}_j^\top \mathbf{W}_j, \end{aligned} \tag{27}$$

where the between-class Laplacian matrix is

$$\mathbf{L}_B = \begin{cases} 2\sum_{p=1}^{C}\sum_{\substack{q=1 \\ p\neq q}}^{C}\left(\frac{V}{N_p^2}\mathbf{e}_p\,\mathbf{e}_p^\top - \frac{1}{N_p N_q}\mathbf{e}_p\,\mathbf{e}_q^\top\right) & \text{if } i = j, \\ -2\sum_{p=1}^{C}\sum_{\substack{q=1 \\ p\neq q}}^{C}\frac{1}{N_p N_q}\mathbf{e}_p\,\mathbf{e}_q^\top & \text{if } i \neq j. \end{cases} \tag{28}$$

$\boldsymbol{m}_p^i$ denotes the mean from the ith view of the pth class in the latent space, and $\mathbf{e}_p$ is the N-dimensional class vector, with N_p as the number of samples in the pth class. The class q is different from the class p.

Alternatively, we propose the between-class scatter matrix which maximizes the distance between different class centers across different views. Since it considers the samples from the class of the specific view origin, we call it Multi-view Modular Discriminant Analysis (MvMDA), and its forumulation is

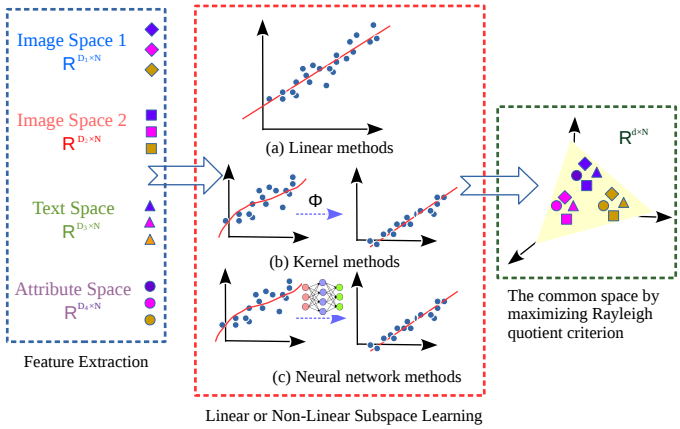


Fig. 3: Overview of the generalized multi-view embedding: Features from different modalities are extracted and either linearly or nonlinearly mapped into the common subspace by maximizing the Rayleigh quotient criterion.

$$\begin{aligned}\mathbf{S}'_B &= \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{p=1}^{C}\sum_{\substack{q=1\\p\neq q}}^{C}(\boldsymbol{m}_p^i-\boldsymbol{m}_q^i)(\boldsymbol{m}_p^j-\boldsymbol{m}_q^j)^\top \\ &= \sum_{i=1}^{V}\sum_{j=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{L}'_B\mathbf{X}_j^\top\mathbf{W}_j, \end{aligned} \tag{29}$$

and the Laplacian matrix is

$$\mathbf{L}'_B = 2\sum_{p=1}^{C}\sum_{q=1}^{C}\left(\frac{1}{N_p^2}\mathbf{e}_p\,\mathbf{e}_p^\top - \frac{1}{N_pN_q}\mathbf{e}_p\,\mathbf{e}_q^\top\right). \tag{30}$$

The difference between the two approaches is that $\mathbf{S}_B$ has $\frac{1}{N_c^2}(V-1)\sum_{i=1}^{V}\sum_{c=1}^{C}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{e}_c\mathbf{e}_c^\top\mathbf{X}_i^\top\mathbf{W}_i$, while $\mathbf{S}'_B$ has the term $\frac{1}{N_c^2}\sum_{i=1}^{V}\sum_{\substack{j=1\\j\neq i}}^{V}\sum_{c=1}^{C}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{e}_c\mathbf{e}_c^\top\mathbf{X}_j^\top\mathbf{W}_j$ which suggests that the first proposal only considers the maximum of the intra-view distances, while the second proposal can maximize the distance between different views. We also validate experimentally that the second proposal achieves better results. Detailed derivation of the two approaches of (27) and (29) are included in the supplementary material.

We extend the same formulation of within-class Laplacian matrix in the latent space as the single-view LDA, i.e.

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\left(\mathbf{I}-\sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top\right)\mathbf{X}_i^\top\mathbf{W}_i \\ &= \sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Q}_{ii}\,\mathbf{W}_i, \end{aligned} \tag{31}$$

where $\mathbf{Q}_{ii} = \mathbf{X}_i\mathbf{L}_W\mathbf{X}_i^\top$, and $\mathbf{L}_W = \mathbf{I} - \sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top$. From (27) and (31), it is shown that the between-class and within-class scatters are equivalent to the projected inter-view and intra-view covariance, respectively. The projection matrix of the multi-view LDA is found by optimizing the following objective function

$$\mathcal{J} = \underset{\mathbf{W}_v,v=1,\ldots,V}{\arg\max}\frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{j=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{L}_B^*\mathbf{X}_j^\top\mathbf{W}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{X}_i\mathbf{L}_W\mathbf{X}_i^\top\mathbf{W}_i\right)}, \tag{32}$$

where $\mathbf{L}_B^*$ is denoted as the Laplacian matrix of either $\mathbf{L}_B$ or $\mathbf{L}'_B$.

C. Kernel-based non-linear subspace learning

Exploiting the kernel trick in (1) and the Representer theorem in (2) and (24) can be expressed as follows

$$\mathbf{Y}_v = \mathbf{A}_v^\top\boldsymbol{\Phi}_v^\top\boldsymbol{\Phi}_v = \mathbf{A}_v^\top\mathbf{K}_v. \tag{33}$$

The criterion of kernel multi-view CCA is then,

$$\mathcal{J} = \underset{\mathbf{K}_v,v=1,\ldots,V}{\arg\max}\frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1\\j\neq i}}^{V}\mathbf{A}_i^\top\mathbf{K}_i\,\mathbf{L}\,\mathbf{K}_j\mathbf{A}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V}\mathbf{A}_i^\top\mathbf{K}_i\,\mathbf{L}\,\mathbf{K}_i\mathbf{A}_i\right)}. \tag{34}$$

It can be easily shown that the solution for $\mathbf{A}_v$ is the same as (19).

Kernel multi-view PLS maximizes the covariance between pairs of feature vectors in the kernel space and therefore the objective function is

$$\mathcal{J} = \underset{K_v,v=1,\ldots,V}{\arg\max}\ \mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1\\j\neq i}}^{V}\mathbf{A}_i^\top\mathbf{K}_i\,\mathbf{L}\,\mathbf{K}_j\mathbf{A}_j\right). \tag{35}$$

The criterion for kernel multi-view discriminant analysis is

$$\mathcal{J} = \underset{K_v,v=1,\ldots,V}{\arg\max}\frac{\mathrm{Tr}\left(\sum_{i=1}^{V}\sum_{j=1}^{V}\mathbf{A}_i^\top\mathbf{K}_i\mathbf{L}_B^*\mathbf{K}_j\mathbf{A}_j\right)}{\mathrm{Tr}\left(\sum_{i=1}^{V}\mathbf{A}_i^\top\mathbf{K}_i\mathbf{L}_W\mathbf{K}_i\mathbf{A}_i\right)} \tag{36}$$

D. Non-linear subspace learning using (deep) neural networks

Exploiting the non-linear mapping using neural networks by (3), (24) can expressed as

$$\mathbf{Y}_v = \mathbf{W}_v^\top h(\mathbf{X}_v; \mathbf{B}_v) = \mathbf{W}_v^\top \mathbf{H}_v. \tag{37}$$

Since the network outputs are combined by a linear layer as shown in Fig. 2, the parameters $\mathbf{B}_v$ of each network are jointly trained to reach the optimal criterion value. After the transformation by neural networks, the projection becomes the same as the multi-view linear subspace learning with respect to $\mathbf{H}_v$. Therefore, we need an additional optimization solved by SGD. We experimented with SGD without variance constraints, and found that we could obtain much better results with the projections constrained to have the unit variance, i.e. in Deep Multi-view CCA (DMvCCA), we have

$$\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{H}_i \, \mathbf{L} \, \mathbf{H}_i^\top \mathbf{W}_i = \mathbf{I}. \tag{38}$$

Without intra-view minimization, the optimization of Deep Multi-view PLS (DMvPLS) is constrained to have unit variance $\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{W}_i = \mathbf{I}$, while in Deep Multi-view Modular Discriminant Analysis (DMvMDA), we project the within-class scatter into unit, i.e.

$$\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{H}_i \mathbf{L}_W \mathbf{H}_i^\top \mathbf{W}_i = \mathbf{I} \tag{39}$$

With the variance constraint, the expressions of the gradients in DMvCCA and DMvPLS are the same as

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{H}_i} &= \frac{\partial}{\partial \mathbf{H}_i} \operatorname{Tr} \left(\sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i^\top \mathbf{H}_i \, \mathbf{L} \, \mathbf{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathbf{H}_j \, \mathbf{L}, \end{aligned} \tag{40}$$

and the gradient of DMvMDA is computed as

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \mathbf{H}_i} &= \frac{\partial}{\partial \mathbf{H}_i} \operatorname{Tr} \left(\sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{H}_i \, \mathbf{L}_B^* \, \mathbf{H}_j^\top \mathbf{W}_j \right) \\ &= \sum_{i=1}^{V} \sum_{j=1}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathbf{H}_j \, \mathbf{L}_B^*, \end{aligned} \tag{41}$$

Detailed derivation of (40) and (41) can be found in the supplementary material.

IV. Experiments

In this section, we evaluate the multi-view methods on two important multimedia applications: zero-shot recognition on the Animal with Attribute (AwA) dataset, and cross-modal image retrieval on the Wikipedia and Microsoft-COCO datasets.

A. Experimental Setup

We conduct the experiments on three popular multimedia datasets. One common property in these datasets is that multi-modal feature representations can be generated. The Animal with Attribute (AwA) dataset consists of 50 animal classes with $30,475$ images in total, and 85 class-level attributes. We follow the same setup as in [31] by splitting 40 classes ($24,295$ images) to train the categorical model while the rest 10 classes with $6,180$ images for testing. Sample images from the test set are shown in Fig. 1. Each animal class contains more than one positive attribute, and the attributes are shared across classes which enables zero-shot recognition. The detailed class labels and attributes are provided in [31].

Wikipedia is a cross-modal dataset collected from the "Wikipedia featured articles" [1]. The dataset is organized in 10 categories and consists of $2,866$ documents. Each document is a short paragraph with a median text length of 200 words, and is associated with a single image. We follow the train/test split in [1] who use $2,173$ training and 693 test pairs of images and documents.

The third dataset we use is the Microsoft COCO 2014 Dataset [47] (abbreviated as COCO in latter paragraphs). We collect the images belonging to at least one fine-grained category, which amounts to $82,081$ training images, and $40,137$ validation images. More than 5 human-annotated different captions are associated to each image. We follow the same definition in [47] to use 12 super classes as the class labels, and 91 fine-grained categories as the attributes. The class names and attributes are presented in Table II. The classes that the images belong to are highly semantic, and the same image can have multiple class labels. Meanwhile, similar images may belong to several different classes.

TABLE II: The class labels and attributes on the COCO dataset.

Classes
outdoor, food, indoor, appliance, sports, person, animal, vehicle, furniture, accessory, electronic, kitchen

Attributes
person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic, light, fire, hydrant, stop, sign, parking, meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports, ball, kite, bat, baseball, glove, skateboard, surfboard, tennis, racket, bottle, wine, glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted, plant, bed, dining, table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy, bear, hair, drier, toothbrush

We use the following feature representations in the experiments:

- **Image feature by CNN models**: We employ the off-the-shelf CNN models as stated in [48] and [**?**] on all image datasets — Visual features are extracted by adopting two powerful pre-trained models. We rescale the size of the input images to 224×224, and generate the features from the outputs of the $fc8$ layer in a VGGNet with 16 weight

layers [49] (denoted as *VGG-16* in latter sections), and the $loss3/classifier$ layer from a GoogleNet [50]. Both models produce 1000-dimension feature vectors.

- **Class label encoding**: Since each image corresponds to one class label on the AwA and Wikipedia dataset, we can describe the image category using the textual feature mapped from the image feature. Specifically, we firstly train a 100-dimension skip-gram model [51] on the entire English Wikipedia articles composed of 2.9 billion words. Then we can extract a separate set of word vectors from class labels of our datasets. In order to correlate the labels with the image contents, we train a ridge regressor with 10-fold cross-validation to map the *VGG-16* image features to each dimension of the word vectors respectively. The regressor outputs are used as the class label features.
- **Attribute encoding**: We also adopt another important modality from visual attributes on the AwA and COCO datasets. On the AwA dataset, we use the 50×85 class-attribute matrix in [52], [53] which specifies attribute probabilities of each class, while on the COCO dataset, we develop a 91-bin feature vector as attributes for each image of which 1's denote the image has the fine-grained tag and 0's otherwise. Then, we train a ridge regressor between the *VGG-16* image feature and formulated attribute probabilities. The predicted probabilities associated with each image are used as the attribute feature.
- **Sentence encoding**: A vital feature of cross-modal retreival system is that we make use of textual features directly. We can find a paragraph of text describing each image on the Wikipedia dataset, while on the COCO dataset, a similar paragraph can be developed by concantenating all captions from the annotators which are associated to each image. We generated the sentence vectors from the paragraphs by the pre-trained skip-thoughts model [54]. The model was trained over the MovieBook and BookCorpus dataset [55]. On the Wikipedia, we employ the *combined-skip* vector of 4800 dimensions, while due to the large size of COCO dataset, we only use the *uni-skip* vector of 2400 dimensions.

The Experiment protocol and performance metrics are described below:

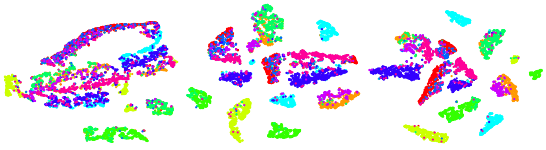
- **Zero-shot recognition on the AwA dataset**: We follow a similar experiment pipeline as in [56], and the comparative results show the performance of the proposed multi-view embedding methods. We project the multi-view representations to the latent space. Zero-shot recognition is achieved by semi-supervised label propagation on a transductive hypergraph in the latent space. Specifically, the cross-domain knowledge learned from the common semantic space is tranferred to the target space of 10 test animal classes via attributes. The prediction of target classes is undertaken on a hypergraph to better integrate different views. We replace the multi-view linear CCA for joint embedding in [56] with the generalized embedding methods. Since the same hypergraph is used, the recognition results indicate the different performance by the multi-view methods in this paper. For the evaluation metric, we use the average classification accuracy which is also employed in [31], [56].
- **Cross-modal retrieval on the Wikipedia and COCO datasets**: We perform two tasks in cross-modal retrieval, i.e. text query for image retrieval and image query for text retrieval. Moreover, a conventional content-based image retrieval system is evaluated in Section IV-C4. We first extract the test features in their own domains. A latent space is joinly learned from the image features, intermediate feature and sentence feature in the training set. Test features are then projected to the latent space by the trained model. The semantic matching from [1] is performed by training a logistic regressor over the embedded features from all of the ground truth samples which maps the projected features of both queries and to-be-retrieved images/texts towards the class labels. The feature vectors generated from the ground truth class labels are essentially the class vectors, whose dimensionality is the number of classes. We use the class probabilities from the regressor outputs for matching between modalities.

  We present the results using 11-point interpolated precision-recall (PR) curves. The Mean Average Precision (MAP) score, which is the average precision at the ranks where recall changes, can be computed based on the Precision Recall curves. The Average Precision (AP) measures the relevance between a query and retrieved items [57], and the MAP score calculates the mean AP by querying all items in the test set.

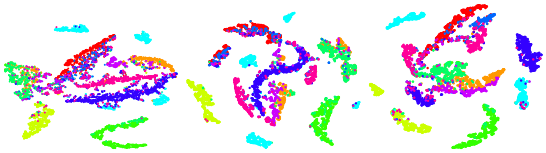
B. Parameter Settings

The dimensionality d in the latent space is a pre-defined parameter. We will evaluate the effects of different d values in the following section. In the experiment, we use $d = 50$ for linear projections on all datasets. On the Wikipedia and AwA dataset, we choose $d = 150$ for kernel mappings, and $d = 200$ for the COCO dataset. For computational efficiency on the AwA and COCO dataset, an approximated RBF kernel mapping is adopted for the non-linear mappings. We set σ in the RBF kernel as the average distance between samples from different views/modalities, which is the natural scaling factor for each dataset. In all of the experiments, the original training set is further partitioned into a 80% training split and a 20% validation split.

The topology of neural networks has more variabilities, and we chose the optimal one according to the held-out validation set. We refer to [58], [59] for a detailed discussion on topologies. On the AwA dataset, we took 3 hidden layer, each with $1,024$ neurons with the *relu* activation before the 50-dimensional linear embedding layer. We only adopted the linear and kernel-based embeddings on the Wikipedia dataset in view of its small size. On the COCO dataset, we chose a single hidden layer with 1500 *relu* neurons, and the dimensionality of the final linear layer is also 1500. We experimented both with the whole batch and multiple mini-batches for SGD, and adopted a batch size of 200 which achieves a superior



(a) 2-view LMvCCA (b) 3-view LMvCCA (c) 4-view LMvCCA



(d) 4-view MvDA [17] (e) 4-view GMA [18] (f) 4-view DMvMDA

Fig. 4: The first row shows the 2-D visualization of embeddings by LMvCCA with an increasing number of views on the AwA dataset. The second row presents the embedding maps by different methods all with 4 views on the same dataset. The samples from different classes are denoted in different colors.

performance. The number of epoches is set to 50 empirically.

C. Experimental Results

The abbreviations of the numerous methods are shown in Table III.

1) Results on zero-shot recogntion: We visualize the embedded space in Fig. 4. We use the VGG-16 feature and class label encoding for two views, and augment attribute and GoogleNet encodings as the additional views. In the first row, it is shown with the increasing number of views in MvCCA, the latent feature vector progresses from being distributed incoherently to showing more distinct groups. In the second row, we compare different methods with 4 views. It is clearly shown we obtain a set of more compact and separable features by the proposed DMvMDA.

Recognition accuracy of different methods is compared quantitatively in Table IV. The first group contains the linear projection results, the second uses the kernel methods, the third are the results by deep neural nets, and the last category includes several comparative results in the literature. The linear methods perform favorably in general while the leading recognition rates can be found in the non-linear methods using neural nets with 4 views. The kernel approximation does not provide superior results compared to linear methods due to the information loss in sampling [28]. Above all, the 4-view DMvMDA is reported to be the best method for zero-shot recognition. The results are also organized by the number of views in columns, and it is shown for all methods that we consistently obtain a better accuracy with more views. Specifically, the proposed LMvPLS achieves the highest accuracy with two input views. while the novel LMvMDA has a more discriminant representation in the latent space leading to a better recognition when more views are presented.

2) Cross-modal retrieval results on the Wikipedia Dataset: Due to the limited number of samples, we use PCA before performing the subspace learning. We use the *VGG-16* and sentence features for two views, and augment attribute and GoogleNet encodings as the additional modalities. It is shown that a better MAP score is obtained when enriching the latent feature with more modalities as shown in Table V. We also observe that the supervised methods perform better than the unsupervised counterparts, and non-linear projections by kernel methods are superior. KMvMDA achieves the best retrieval results with supervision and non-linearity.

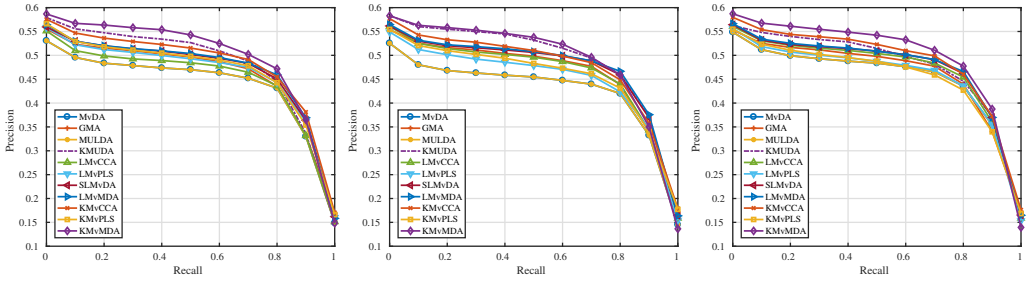
We present more detailed results in the form of PR curves in Fig. 5. For image queries, KMvMDA consistently outperforms the other methods across all views, which can be explained by its utilization of class labels and kernel-based representations. For text queries, the supervised and non-linear methods also outperform their linear counterparts. KMvCCA and KMvMDA are the leading methods in this category, which shows the strength of cross-modal retrieval by making use of view difference.

3) Cross-modal retrieval results on the COCO Dataset: The COCO dataset is much larger than the Wikipedia dataset, and we pay more attention to the non-linear methods especially the ones using neural networks. Many images have more than one class labels, and therefore we focus on the unsupervised learning algorithms. Similar to the experiments above, the MAP scores in Table VI show that a gain of retrieval accuracy can be obtained by embedding additional modalities into the latent space. DCCA2 [25] achieves a superior performance with 2 views thanks to its non-linear projection which makes the latent feature more discriminant for retrieval. However, its formulation limits the algorithm to 2 views, and DMvCCA and DMvPLS based on the proposed framework can improve the state-of-the-art method by increasing the number of modalities. From the PR curves in Fig. 6, we compare the methods using the proposed objective function with DCCA2 which contains two views. For image queries, KapMvCCA obtains the best retrieval result with 2 views, but it is further improved by the methods using neural networks benefitted by attributes and GoogleNet features. For text queries, it also suggests more modalities and neural network-based representations contribute to the retrieval performance. The cross-modal retrieval by the 4-view DMvCCA achieves the overall highest precision score on this dataset.

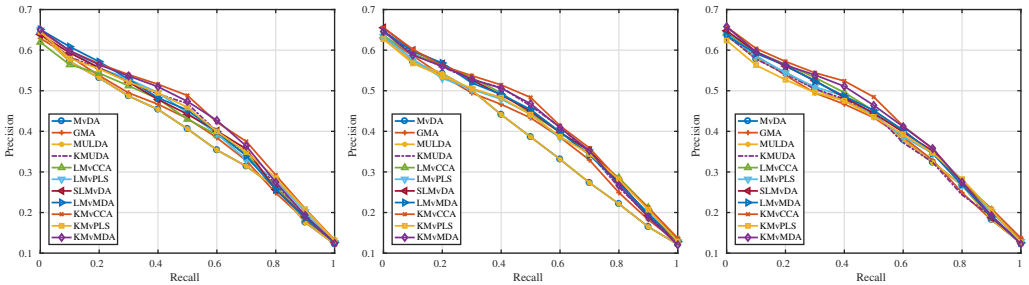
4) Content-based Image Retrieval (CBIR) Performance on the COCO dataset: We also show the effectiveness of multi-view embedding method on the conventional CBIR task in Fig. 7. We randomly pick two image-to-text pairs as queries, to perform image-to-image retrieval using both the *VGG-16* visual feature and the projected visual feature by the 4-view DCCA. We also perform text-to-image retrieval by querying the corresponding captions of the query image used in CBIR in the last column. We observe the CBIR performance can be further improved by incorporating the semantic information. In Table VII, we present the quantitative results of CBIR by the projected visual features. "RAW" in the Table shows the retrieval results by visual features directly, while the rest are the multi-view embedding results. It is shown that more modalities and non-linear projections yield a discriminant latent visual feature, which improves the retrieval performance.

TABLE III: List of Abbreviations

LMvCCA / KMvCCA / KapMvCCA / DMvCCA	Linear / Kernel / Approximate Kernel / Deep Multi-view Canonical Correlation Analysis
LMvPLS / KMvPLS / KapMvPLS / DMvPLS	Linear / Kernel / Approximate Kernel / Deep Multi-view Partial Least Square Regression
SLMvDA/ SKMvDA	Standard Linear / Kernel Multi-view Discriminant Analysis using (28)
LMvMDA / KMvMDA / KapMvMDA / DMvMDA	Linear / Kernel / Approximate Kernel / Deep Multi-view Modular Discriminant Analysis using (30)
MULDA / KMUDA [15]	Multi-view Uncorrelated Linear / Kernel Discrimiant Analysis
MvDA [17]	Multi-view Discrimiant Analysis
GMA [18]	Generalized Mult-view Analysis
DCCA2 [25]	Deep Canonical Correlation Analysis



(a) PR curves for image queries with 2 views (b) PR curves for image queries with 3 views (c) PR curves for image queries with 4 views



(d) PR curves for text queries with 2 views (e) PR curves for text queries with 3 views (f) PR curves for text queries with 4 views

Fig. 5: PR curves across different number of views on the Wikipedia dataset for the Image-to-Text retrieval and the Text-to-Image retrieval.

TABLE IV: RECOGNITION ACCURACY (%) on the AwA DATASET

Method	2 views	3 views	4 views
Proposed LMvCCA	55.86	75.88	82.01
Proposed LMvPLS	**58.52**	73.59	77.09
Proposed LMvMDA	55.85	**77.64**	**82.88**
Proposed SLMvDA	54.58	69.02	70.56
Proposed KapMvCCA	56.41	73.40	74.76
Proposed KapMvPLS	55.58	**74.40**	75.05
Proposed KapMvMDA	**57.19**	71.64	**75.63**
Proposed DMvCCA	51.25	71.12	82.27
Proposed DMvPLS	43.28	68.81	74.63
Proposed DMvMDA	**53.87**	**75.61**	**83.66**
MvDA [17]	49.95	68.55	70.00
GMA [18]	52.12	73.49	78.46
MULDA [15]	55.46	74.13	74.88
TMV-HLP [56]	-	73.50	80.50
DCCA2 [25]	50.47	-	-

D. Parameter sensitivity analysis of dimension d in linear and kernel cases

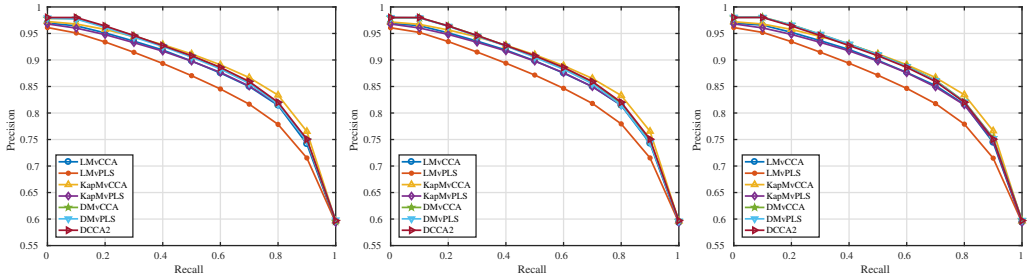
The number of dimension of the feature vectors in the latent space is determined by the top d eigenvectors in the projection matrix, and it is pre-defined in the former experiments. Therefore in this section, we investigate the effect by the variation of d shown in Fig. 8 and 9, ranging from $\{10, 20, 50, 100, 150, 200\}$. The performance on the Wikipedia dataset is reported with both text queries on images and image queries on texts. The results on different number of views are also recorded. In general, we obtain a better retrieval performance when d is between 50 and 150. It can be explained by the fact that the most informative eigenvectors are included within the range. Therefore, $d = 50$ was chosen for the multi-view linear embeddings in the experiments. Except LMvPLS and KMvPLS, we find the majority of the methods are robust to the dimensionality changes in the subspace.

TABLE V: MAP Scores (%) on the Wikipedia

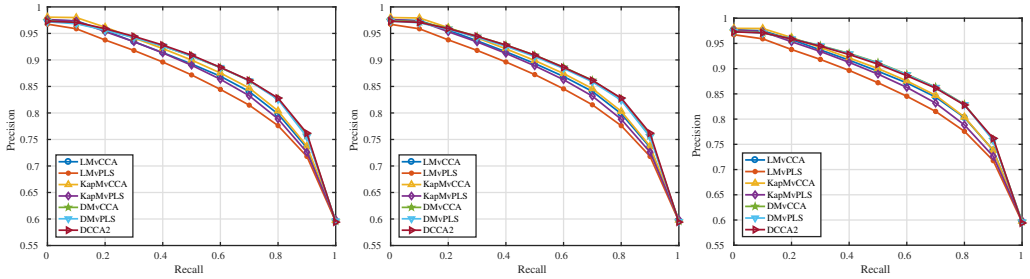
	2 views			3 views			4 views		
	img. query	txt. query	avg.	img. query	txt. query	avg.	img. query	txt. query	avg.
MvDA [17]	39.73	37.14	38.43	39.34	35.04	37.19	41.07	39.21	40.14
GMA [18]	41.91	38.55	40.23	42.26	38.66	40.46	42.26	38.67	40.47
MULDA [15]	43.04	39.87	41.46	43.45	40.68	42.07	43.79	40.32	42.06
Proposed LMvCCA	41.37	39.07	40.22	42.10	39.64	40.87	42.53	39.98	41.26
Proposed LMvPLS	42.49	40.42	41.46	41.29	39.34	40.31	41.86	39.74	40.80
Proposed SLMvDA	43.20	40.07	41.64	43.14	39.86	41.50	43.77	40.24	41.80
Proposed LMvMDA	43.38	40.32	**41.85**	43.74	40.46	**42.10**	43.90	40.23	**42.07**
KMUDA [15]	44.38	39.52	41.95	45.40	39.96	42.68	44.29	38.12	41.20
Proposed KMvCCA	44.78	41.83	43.30	44.06	41.41	42.73	45.13	41.66	43.40
Proposed KMvPLS	42.94	40.46	41.70	42.03	39.40	40.71	41.94	38.84	40.39
Proposed SKMvDA	45.52	38.39	41.96	44.66	38.47	41.57	42.94	39.32	41.13
Proposed KMvMDA	46.01	40.96	**43.49**	45.40	40.16	**42.78**	46.48	40.73	**43.61**

TABLE VI: MAP Scores (%) on the COCO dataset

	2 views			3 views			4 views		
	img. query	txt. query	avg.	img. query	txt. query	avg.	img. query	txt. query	avg.
Proposed LMvCCA	87.18	86.92	87.05	87.20	87.01	87.11	87.31	87.22	87.27
Proposed LMvPLS	84.76	85.05	84.91	84.83	85.07	84.95	84.82	85.05	84.94
Proposed KapMvCCA	88.42	87.58	88.00	88.35	87.52	87.94	88.45	87.60	88.03
Proposed KapMvPLS	87.16	86.58	86.87	87.14	86.56	86.85	87.14	86.56	86.85
Proposed DMvCCA	88.14	88.10	88.12	88.20	88.26	**88.23**	88.49	88.40	**88.45**
Proposed DMvPLS	88.01	88.03	88.02	88.06	88.03	88.05	88.45	88.34	88.40
DCCA2 [25]	88.30	88.27	**88.29**	-	-	-	-	-	-



(a) PR curves for image queries with 2 views (b) PR curves for image queries with 3 views (c) PR curves for image queries with 4 views



(d) PR curves for text queries with 2 views (e) PR curves for text queries with 3 views (f) PR curves for text queries with 4 views

Fig. 6: PR curves across different number of views on the COCO dataset for the Image-to-Text retrieval and the Text-to-Image retrieval. Note the curve by DCCA2 [25] is presented across all numbers of views.

V. Conclusion

In this paper, we proposed a generalized multi-view embedding method using the graph embedding framework. We showed multi-view CCA, PLS and LDA can be characterized by their specific intrinsic and penalty graph matrices within the same framework. A novel discriminant analysis method named MvMDA was introduced by exploiting the distances between class centers of different views. Meanwhile, we also

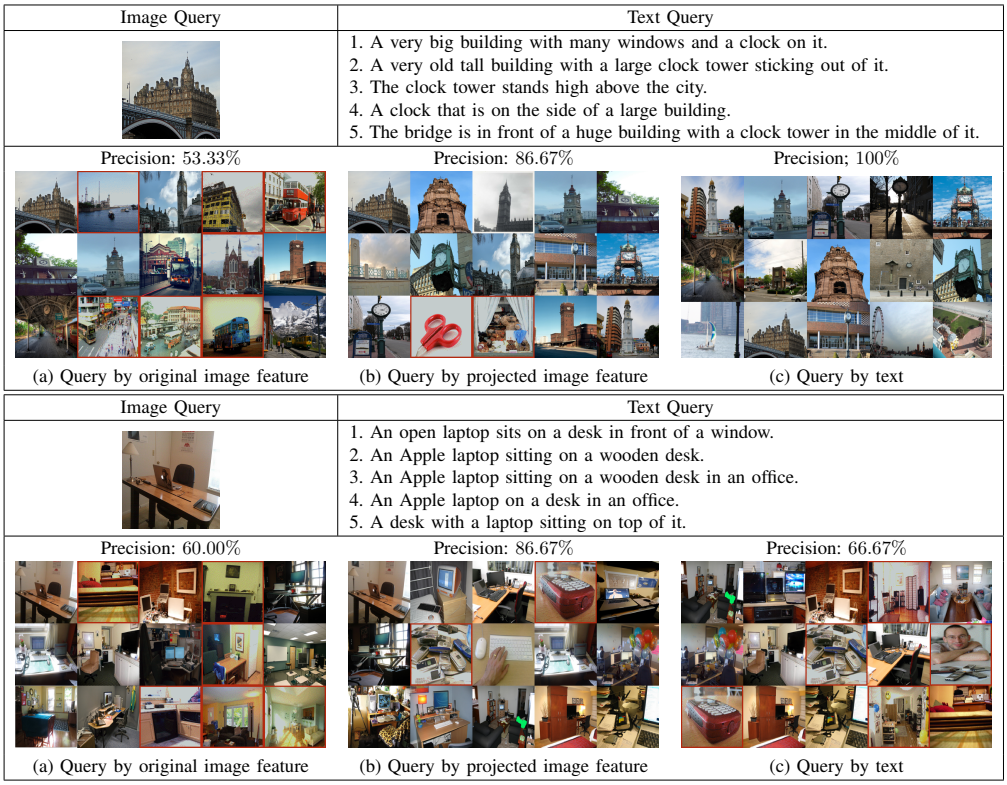
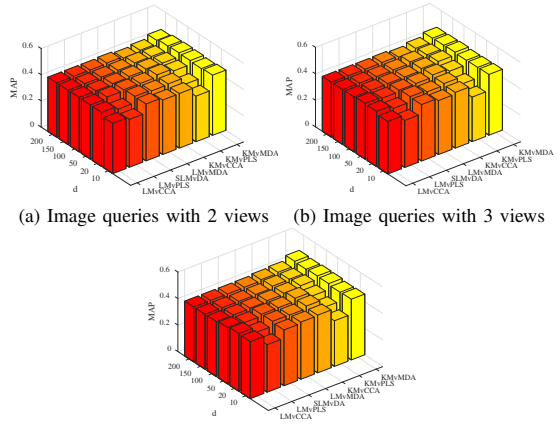


Fig. 7: Sample retrieval results on the COCO dataset. The first row of each table presents the query image and text, and the second row shows the retrieved images by different query types. False positive results are bounded in red.

TABLE VII: MAP(%) scores of CBIR on the COCO dataset

Method	2 views	3 views	4 views
Raw		83.77	
Proposed LMvCCA	85.64	85.76	85.93
Proposed LMvPLS	84.30	84.30	84.32
Proposed KapMvCCA	85.43	85.47	85.49
Proposed KapMvPLS	84.56	84.57	84.58
Proposed DMvCCA	89.33	**89.62**	**89.84**
Proposed DMvPLS	89.50	89.34	89.79
DCCA2 [25]	**89.71**	-	-



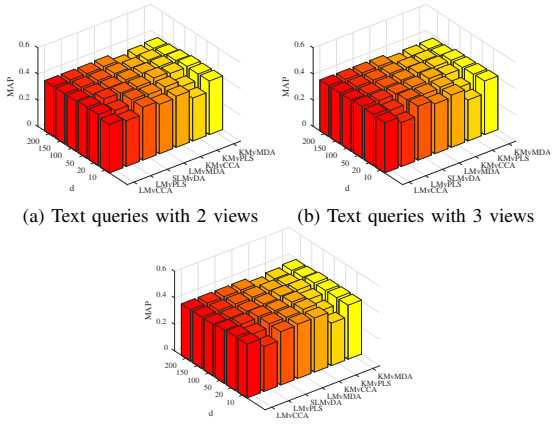
(a) Image queries with 2 views (b) Image queries with 3 views

(c) Image queries with 4 views

Fig. 8: Performance variation for image queries on texts of Wikipedia dataset with respect to the different dimension d.

studied non-linear embeddings, and found implicit and explicit kernel mappings for multi-view learning. A unified scheme for learning by neural networks was developed which combined the learned representations with a linear embedding layer. We thereby formulated the expression of stochastic gradient descent for optimizing the proposed objective function.

We validated the formulation by conducting experiments in zero-shot visual object recognition and cross-modal image retrieval. It was shown that supervised and non-linear subspace learning outperformed the unsupervised and linear methods when large amount of images and texts were available. Moreover, the recognition or retrieval performance were consistently improved by embedding more views/modalities into the latent feature space. We also performed the traditional CBIR experiments where the multi-view embeddings can contribute



(a) Text queries with 2 views (b) Text queries with 3 views

(c) Text queries with 4 views

Fig. 9: Performance variation for text queries on images of Wikipedia dataset with respect to the different dimension d.

to the performance gain.

Interesting future research directions include learning from the raw data to achieve an end-to-end solution for multi-view learning. We should further reduce the computational cost for kernel methods to cope with large scale of images. In addition, learning from incomplete and unlabeled multi-view data should be studied for video analysis.

REFERENCES

[1] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.

[2] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[3] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate analysis*. Academic press, 1980, ch. 10 Canonical Correlation Analysis, pp. 281–290.

[4] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing (TIP)*, vol. 11, no. 3, pp. 293–305, 2002.

[5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[7] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, Nov 2015.

[8] S. Wold, A. Ruhe, H. Wold, and W. Dunn, III, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[9] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Multiview fisher discriminant analysis," in *NIPS workshop on learning from multiple sources*, 2008.

[10] G. Cao, M. A. Waris, A. Iosifidis, and M. Gabbouj, "Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Dec 2016, pp. 1–6.

[11] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[12] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, Sept 2013.

[13] T. Sun, S. Chen, J. Yang, and P. Shi, "A novel method of combined feature extraction for recognition," in *In Proceedings of IEEE International Conference on Data Mining, (ICDM)*. IEEE, 2008, pp. 1043–1048.

[14] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," in *Proceedings of the 24th international conference on Machine learning (ICML)*. ACM, 2007, pp. 577–584.

[15] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3272–3284, Dec 2016.

[16] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1123–1128.

[17] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[18] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[19] J. Liu, Y. Jiang, Z. Li, Z. H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1233–1246, June 2015.

[20] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 8, pp. 1559–1572, Aug 2014.

[21] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[22] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[23] T. Sun, S. Chen, Z. Jin, and J. Yang, "Kernelized discriminative canonical correlation analysis," in *International Conference on Wavelet Analysis and Pattern Recognition,*, vol. 3. IEEE, 2007, pp. 1283–1287.

[24] A. Iosifidis and M. Gabbouj, "Scaling up class-specific kernel discriminant analysis for large-scale face verification," *IEEE Transactions on Information Forensics and Security*, vol. PP, no. 99, pp. 1–1, 2016.

[25] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1247–1255.

[26] M. Dorfer, R. Kelz, and G. Widmer, "Deep linear discriminant analysis," *International Conference on Learning Representations (ICLR)*, 2016.

[27] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 12, pp. 5017–5032, Dec 2015.

[28] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[29] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1778–1785.

[30] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 951–958.

[31] ——, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.

[32] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, March 2015.

[33] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-modal subspace learning via pairwise constraints," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5543–5556, Dec 2015.

[34] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2431–2442, Dec 2014.

[35] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, May 2014.

[36] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE Transactions on Cybernetics*, 2017, to be published.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[38] T. Hastie, R. Tibshirani, and J. Friedmani, *The Elements of Statistical Learning*. Springer, 2013, ch. 18. High-Dimensional Problems:$p \gg N$, pp. 649–694.

[39] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[40] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of Annual Conference of Computational Learning Theory,*. Springer, Heidelberg, Germany, 2001, pp. 416–426.

[41] M. Borga, "Canonical correlation: a tutorial," http://people.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf, 2001.

[42] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4590–4594.

[43] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 23, no. 2, pp. 228–233, 2001.

[44] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[45] A. Iosifidis, A. Tefas, and I. Pitas, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

[46] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. [Online]. Available: http://mscoco.org/home/

[48] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2015, pp. 1–9.

[51] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.

[52] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, vol. 3, 2006, p. 5.

[53] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith, "Default probability," *Cognitive Science*, vol. 15, no. 2, pp. 251–269, 1991.

[54] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.

[55] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," *Cai2007*, 2015.

[56] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.

[57] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to Information Retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1, ch. 8. Evaluation in information retrieval, pp. 188–210.

[58] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423–1447, 1999.

[59] S. Kiranyaz, T. Ince, A. Yildirim, and M. Gabbouj, "Evolutionary artificial neural networks by multi-dimensional particle swarm optimization," *Neural Networks*, vol. 22, no. 10, pp. 1448 – 1462, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6T08-4WGF117-1/2/600b51dc41c51f5fc4c9427b352c7e6a

Guanqun Cao received the double B.Eng. degree in Electronic and Information/Computer Engineering from Huazhong University of Science and Technology, China and University of Birmingham, UK. He also received the M.Sc degree from the joint Erasmus Mundus programme in Color Informatics and Media Technology. He is currently a PhD student at the Multimedia Research Group, Tampere University of Technology, Finland. His research interests include multimedia retrieval and machine learning with a focus on multi-view data analysis.

Alexandros Iosifidis (SM'16) received the Diploma and the M.Eng. degrees in Electrical & Computer Engineering from Democritus University of Thrace in 2008 and 2010, respectively. He received the Ph.D. in Informatics from Aristotle University of Thessaloniki in 2014. Currently, he is an Assistant Professor in the Department of Engineering, Electrical and Computer Engineering, Aarhus University, Denmark. Before that, he was a postdoctoral researcher at the Multimedia Research Group of the Department of Signal Processing in Tampere University of Technology, holding an Academy Postdoctoral Research Fellow position.

Dr. Iosifidis is a Senior member of IEEE. He has (co-)authored more than 100 journal and conference papers in his areas of expertise. His research interests are in the areas of pattern recognition and machine learning, with applications mainly in images/videos and time series.

Ke Chen was born in Wuxi, China in 1985. He received Ph.D major in computer vision under the supervision of Prof. Shaogang Gong and Prof. Tao Xiang at School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK. He received his B.E. major in automation and M.E. major in software engineering supervised by Prof. Yunong Zhang at Sun Yat-sen University, China in 2007 and 2009, respectively.

Dr. Chen is currently the Academy of Finland post-doctoral research fellow at the Department of Signal Processing, Tampere University of Technology. His research interests include computer vision, pattern recognition, neural dynamic modelling, and robotic inverse kinematics. He has published more than forty peer-reviewed conference and journal papers in computer vision, neural networks and robotics.

Moncef Gabbouj (F'11) received his BS degree in electrical engineering in 1985 from Oklahoma State University, and his MS and PhD degrees in electrical engineering from Purdue University, in 1986 and 1989, respectively. Dr. Gabbouj is a Professor of Signal Processing at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Academy of Finland Professor during 2011-2015. His research interests include multimedia content-based analysis, indexing and retrieval, machine learning, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding. Dr. Gabbouj is a Fellow of the IEEE and member of the Academia Europaea and the Finnish Academy of Science and Letters. He is the past Chairman of the IEEE CAS TC on DSP and committee member of the IEEE Fourier Award for Signal Processing. He served as associate editor and guest editor of many IEEE, and international journals and Distinguished Lecturer for the IEEE CASS. He organized several tutorials and special sessions for major IEEE conferences and EUSIPCO. Dr. Gabbouj guided 46 PhD students and published 700 papers.

Publication II

G. Cao, A. Iosifidis and M. Gabbouj, “Multi-View Nonparametric Discriminant Analysis for Image Retrieval and Recognition,” in IEEE Signal Processing Letters, vol. 24, no. 10, pp. 1537-1541, Oct. 2017. doi: 10.1109/LSP.2017.2748392

Multi-view Nonparametric Discriminant Analysis for Image Retrieval and Recognition

Guanqun Cao, Alexandros Iosifidis, *Senior Member, IEEE*, Moncef Gabbouj, *Fellow, IEEE*

***Abstract*—A novel multi-view nonparametric discriminant analysis method is proposed for the application of cross-modal image retrieval and zero-shot recognition. We exploit the class boundary structure and discrepancy information of the available views in order to formulate an optimization criterion which is automatically adjusted to the multi-view class structures. The proposed method allows for multiple projection directions, by relaxing the Gaussian distribution assumption of related methods. The experiments demonstrate that the proposed method can achieve superior results comparing to several existing methods.**

***Index Terms*—Multi-view learning, subspace learning, image retrieval**

I. Introduction

We have entered a world of multimedia big data. Multimedia contents also become increasingly diverse in their representation and exist in different modalities. It urges the research community to dive into the heterogeneous data to find the desired content across modalities or classify them into the right category from many views. For example, thanks to the available text-image datasets from the collaborative content creations in Wikipedia, matching textual description with their corresponding images becomes a hot-button issue. People start to revisit the image retrieval problem not only in the conventional way of retrieving the best matching image using the query text, but generating human understandable sentences given an image [1]. A visual object can also be observed in various domains in terms of illumination, noise level, viewing angle, and self deformation. Integrating the knowledge obtained from multiple views/modalities contributes to improving the task of object recognition [2].

Subspace learning has proved to be successful among the techniques in multi-view learning for multimedia analysis [3]. It finds a common latent space from different input modalities by fitting an optimization criterion. Among unsupervised methods, Canonical Correlation Analysis (CCA) [4] has been widely used to establish a correlation between views [5], [6]. On the other hand, Multi-view Discriminant Analysis (MvDA) [2] as a supervised algorithm is a direct extension of Linear Discriminant Analysis (LDA) [7], [8]. It seeks for the most discriminant features by maximizing the determinant of the between-class scatters while minimizing

The authors are with the Laboratory of Signal Processing, Tampere University of Technology, Finland. A. Iosifidis is also with the Dept. of Engineering, Electrical and Computer Engineering, Aarhus University, DK-8200, Aarhus N, Denmark.

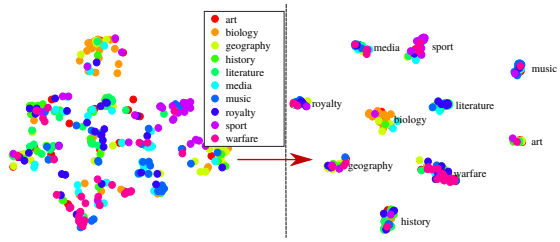


Fig. 1: t-SNE visualization of word2vec represenations before and after applying the proposed method. The samples are grouped together automatically, and each class label indicates the majority class in its group, which matches the corresponding test class.

that of the within-class scatters regardless of view origins. This method can be further extented to nonlinear cases by using (approximate) kernel mappings [9], [10], or integrating with neural nets [11], [6]. Generalized Multi-view Analysis (GMA) [12] was proposed as a framework for numerous techniques to maximize the intra-view discriminant information.

MvDA has certain limitations originating from LDA [13], which is developed upon the assumption that data in each class follow a Gaussian distribution. Only class centers are considered when calculating the between-class scatter matrix and within-class matrix. These parametric methods also suffer performance degradation when the data is non-Gaussian. Several nonparametric techniques [14], [15] were thereby developed to design alternative between-class scatters by exploiting the distances of the data close to the class boundary. However, these techniques are applied in the single-view cases, and view discrepancies should not be overlooked using direct extensions in the multi-view learning.

We propose a new formulation for multi-view discriminant analysis which successfully exploits the boundary structure of the classes on data from different sources, as well as the view discrepancy for balancing the contribution of each view in the overall optimization process. Following the graph embedding framework [16], we design the intrinsic and penalty graphs characterizing the within-class compactness and between-class separability, while encoding both intra-view and inter-view discrimination simultaneously. Class compactness is encoded using a k_1-nearest neighbor graph connecting neighboring samples from the same class with the same view origin, while class discrimination is modeled using another k_2-nearest neighbor graph connecting nearest sample pairs from the same

view but belonging to different classes. We also enhance the class discrimination of each node in the penalty graph by weighting the contribution of neighboring pairs based on their proximity to the class boundary. Moreover, global class discrimination is combined to the adaptive local graph to better adjust to the properties of heterogeneous classes.

We outline the strength of the proposed method as follows: 1) It allows for a larger number of projection directions than MvDA, and makes use of all the samples when developing the intrinsic and penalty graphs, while MvDA merely uses the class centers. 2) It assumes that each class is formed by multiple subclasses, denoted by the different views. In this way, it relaxes the assumption of MvDA in that each class is formed by samples drawn from a multi-dimensional Gaussian distribution, independent from the view they come from. 3) By exploiting both the between-class and within-class margins in the same view, we obtain a better class discrimination in the penalty graph and compactness in the intrinsic graph, and result in an improved performance. 4) Multi-view extension of Marginal Fisher Analysis (MFA) under the GMA framework [12] only considers the intra-view discriminant information, while MvNDA also takes into account of the inter-view discrimination.

The rest of the letter is organized as follows. In Section II, we will present our multi-view nonparametric discriminant analysis in detailed after describing the previous work on MvDA [2]. In Section III, we present quantitative results in cross-modal image retrieval on the Wikipedia dataset and zero-shot recognition on the Animal with Attribute (AwA) dataset. Finally, Section IV concludes the letter.

II. Approach

We denote the data matrix by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. In the multi-view case, we define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}, v = 1, \ldots, V$ for the feature vectors of the vth view. The dimensionality of the various feature spaces D_v can vary across the views. $\mathbf{W} = [\mathbf{W}_1^\top, \mathbf{W}_2^\top, \ldots, \mathbf{W}_V^\top]^\top$, where $\mathbf{W}_v \in \mathbb{R}^{D_v \times d}, v = 1, \ldots, V$ is the projection matrix in view v, d is the number of dimensions in the latent (common) space. For multi-class learning problems, the class label of the sample $\mathbf{x}_i$ is defined as $c_i \in \{1, 2, \ldots, C\}$, where C is the number of classes. We also denote the index set of the cth class by π_c.

We use the graph embedding notation, where we define by $\mathbf{G} = \{\mathbf{X}, \mathbf{V}\}$ an undirected weighted graph with vertex set $\mathbf{X}$ and simlarity matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$. The diagonal matrix $\mathbf{D}$ and the Laplacian matrix $\mathbf{L}$ of a graph $\mathbf{G}$ in the vth view are denoted as $\mathbf{L}_v = \mathbf{D}_v - \mathbf{V}_v$, $\mathbf{D}_{ii}^v = \sum_{j \neq i} \mathbf{V}_{ij}^v$, $\forall i$.

A. Multi-view Discriminant Analysis (MvDA)

MvDA [2] is the multi-view version of parametric LDA which maximizes the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. Mathematically, it is written as

$$\mathcal{J}_{\text{MvDA}}(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^P)}{\text{Tr}(\mathbf{S}_W^P)}, \tag{1}$$

where the between-class scatter matrix is

$$\mathbf{S}_B^P = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \underbrace{\left(\sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top - \frac{1}{N} \mathbf{e}\,\mathbf{e}^\top \right)}_{\mathbf{L}_B^P} \mathbf{X}_j^\top \mathbf{W}_j \tag{2}$$

and the within-class scatter matrix is

$$\mathbf{S}_W^P = \sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i \underbrace{\left(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c} \mathbf{e}_c \mathbf{e}_c{}^\top \right)}_{\mathbf{L}_W^P} \mathbf{X}_j^\top \mathbf{W}_j \tag{3}$$

$\mathbf{L}_B^P$ and $\mathbf{L}_W^P$ are the between-class Laplacian matrix and within-class Laplacian matrix, respectively [17]. Both the single-view and multi-view linear discriminant analysis are parametric methods under the assumption that the data of each class follows a Gaussian distribution. Their performance degrades when the data distribution is non-Gaussian. Moreover, since the rank of the between-class matrix is at most $C-1$ in the vth view, the number of the final MvDA feature is at most $(C-1) \times V$. The classification performance is constrained by the limited number of dimensionality in the subspace.

B. Proposed Multi-view Nonparametric Discriminant Analysis (MvNDA)

We propose a new criterion to learn a mapping from the multiple feature spaces defined over the various views to a common space as follows,

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{S}_B^N)}{\text{Tr}(\mathbf{S}_W^N)}, \tag{4}$$

where $\mathbf{W}$ is the projection matrix containing the eigenvectors of $\mathbf{S} = {\mathbf{S}_W^N}^{-1}\mathbf{S}_B^N$ associated with the top d eigenvalues λ, and can be solved efficiently from the generalized eigenvalue problem as in [2], [6]. We define the within-class scatter matrix $\mathbf{S}_W^N$ and between-class scatter matrix $\mathbf{S}_B^N$ as follows. In the latent space, we enforce the samples from the same class of the same view to be close to each other. Therefore, the intrinsic graph is designed to strengthen the intra-view class compactness from these subclasses, and the within-class scatter matrix is

$$\mathbf{S}_W^N = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i (\mathbf{D}_W - \mathbf{V}_W) \mathbf{X}_i^\top \mathbf{W}_i \tag{5}$$

where $\mathbf{L}_W^N = \mathbf{D}_W - \mathbf{V}_W$ is the within-class Laplacian matrix and the intrinsic graph $\mathbf{V}_W$ is defined as

$$\mathbf{V}_{pq}^W = \begin{cases} 1, & \text{if } p \in \text{NN}_{k_1}(q) \text{ or } q \in \text{NN}_{k_1}(p) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$\text{NN}_{k_1}(p)$ denotes the index set of the k_1 nearest neighbors of the sample $\mathbf{x}_p$ in the same class.

We also design a view-specific penalty graph to push apart the marginal samples from different classes of the same view with the following between-class scatter matrix:

$$\mathbf{S}_B^{\text{VS}} = \sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{X}_i [\mathbf{Q} \circ (\mathbf{D}_B - \mathbf{V}_B)] \mathbf{X}_i^\top \mathbf{W}_i, \tag{7}$$

where $\mathbf{L}_B^{\text{VS}} = \mathbf{D}_B - \mathbf{V}_B$ is the between-class view-specific

Laplacian matrix, and its intrinsic graph is characterized as:

$$\mathbf{V}^B_{pq} = \begin{cases} 1, & \text{if } (p,q) \in \text{NP}_{k_2}(c_p) \text{ or } (p,q) \in \text{NP}_{k_2}(c_q) \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

$\text{NP}_{k_2}(c)$ is a set of data pairs which contains the k_2 nearest pairs in the set $\{(i,j), i \in \pi_c, j \notin \pi_c\}$. The weight matrix $\mathbf{Q}$ aims to highlight the importance of the samples on the classification boundary. Specifically, the value in $\mathbf{Q}$ goes to 0.5 if the sample falls close to the boundary, but reduces to 0 otherwise. $d(p,q)$ is the Euclidean distance between two vectors p and q. $\mathbf{Q}$ is mathematically described below,

$$\mathbf{Q}_{pq} = \begin{cases} \frac{\min\{d(p,q), d(p,\text{NN}_{k_2}(p))\}}{d(p,q)+d(p,\text{NN}_{k_2}(p))} & \text{if } (p,q) \in \text{NP}_{k_2}(c_p) \\ & \text{or } (p,q) \in \text{NP}_{k_2}(c_q) \\ 0 & \text{otherwise.} \end{cases}$$

In order to enforce both inter-view and intra-view class discrimination, our penalty term is based on the linear combination of $\mathbf{S}^P_B$ of MvDA (2) and $\mathbf{S}^{\text{VS}}_B$ of (7) as follows

$$\mathbf{S}^N_B = \alpha\, \mathbf{S}^P_B + (1-\alpha)\, \mathbf{S}^{\text{VS}}_B, \quad (9)$$

where $\alpha \in [0,1]$ is a weighting factor which is set close to 1 if the training data has a Gaussian distribution, and some other value if the data distribution is unknown.

We provide a qualitative illustration of the intrinsic and penalty graph in Fig. 2. The intrinsic graph shows the within-class compactness by connecting a sample to its k_1-nearest-neighbors of the same class and view. The between-class separability is characterized by both the connected marginal point pairs from the same view but of different classes, and the distance of different class centers.

We also follow the standard kernel-based learning approach to define non-linear multi-view mappings. Each input space is then mapped to the so-called kernel space $\mathcal{F}_v$ using a non-linear function ϕ, i.e. $\mathbf{X}_v \in \mathbb{R}^{D_v \times N} \xrightarrow{\mathbf{\Phi}(\cdot)} \mathbf{\Phi}(\mathbf{X}_v) \in \mathbb{R}^{|\mathcal{F}_v| \times N}$. In $\mathcal{F}_v$, following the Representer Theorem [18], [19], a linear projection can be expressed as $\mathbf{W}_v = \mathbf{\Phi}(\mathbf{X}_v)\mathbf{A}_v$ and dot products between data pairs can be expressed using the kernel matrix $\mathbf{K}_v = \mathbf{\Phi}(\mathbf{X}_v)^\top \mathbf{\Phi}(\mathbf{X}_v)$ [20]. Then,

$$\mathcal{J}_{\text{MvNDA}}(\mathbf{A}) = \arg\max_{\mathbf{A}} \frac{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}^N_B\, \mathbf{K}\mathbf{A})}{\text{Tr}(\mathbf{A}^\top \mathbf{K}\, \mathbf{L}^N_W\, \mathbf{K}\mathbf{A})}, \quad (10)$$

where the between-class Laplacian matrix $\mathbf{L}^N_B = \alpha \mathbf{L}^P_B + (1-\alpha)\mathbf{L}^{\text{VS}}_B$, and $\mathbf{K} = \text{diag}(\mathbf{K}_1, \cdots, \mathbf{K}_V)$. For the cases where the direct solution of (10) is impractical, due to the training data size, we employ the approximate kernel mapping proposed in [10] followed by the linear mapping defined in (4).

III. Experiments

A. Wikipedia dataset

The cross-modal retrieval dataset named "Wikipedia" was collected from the "Wikipedia featured articles" [1]. The dataset has 10 generic classes and is composed of $2,866$ documents. Each document is a short paragraph with a median text length of 200 words, and is coupled with a single image.

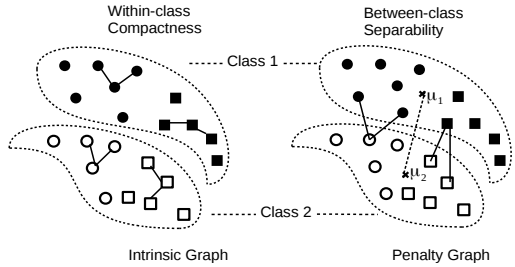


Fig. 2: The adjacency relationship of the intrinsic and penalty graphs of the proposed MvNDA. The circular and rectangle dots indicate samples from different views. We illustrate the 2-nearest adjacencies (i.e. $k_1 = k_2 = 2$) of one sample in each class per view origin for clarity.

We follow the train/test split in [1] using $2,173$ training and 693 test pairs of images and documents. Furthermore, a validation set is held out by 20% of the training image/text pairs. We perform PCA beforehand and reduce the dimensionality of input features to 100. We set the dimensionality of the latent space d to 50 for all methods, and the maximal number of dimesional by MvDA is used. We set $\alpha = 0.5$ and $k_1 = k_2 = 20$ in all experiments based on the validation set.

Here, we briefly describe the feaures extracted from each view in this dataset. For images, two off-the-shelf CNNs models are used to produce the visual features. VGGNet provides the *view 1* feature using the output from the $fc8$ layer in VGGNet with 16 weight layers [21]. We also use the GoogleNet outputs as the *view 3* features. *View 2* feature is extracted from the Wikipedia paragraphs surrounding the images using a pre-trained skip-thoughts model [22]. An additional *view 4* feature is the regression outputs from the Word2Vec by mapping the visual feature to the word feature [23]. The same set of features has been adopted and detailed description can be found in [6].

The cross-modal retrieval is conducted in both ways by

TABLE I: MAP Score (%) on the Wikipedia Dataset

	Linear methods			Kernel methods		
Method	img. query	txt. query	Avg.	img. query	txt. query	Avg.
2 views						
MvCCA [6]	36.92	34.96	35.94	44.78	41.83	43.31
MvPLS [6]	42.49	40.42	41.46	42.94	40.46	41.70
GMA [12]	41.91	38.55	40.23	45.65	36.97	41.31
MvDA [2]	39.73	37.14	38.44	44.16	37.82	40.99
MvNDA	43.51	40.72	**42.12**	48.41	41.97	**45.19**
3 views						
MvCCA [6]	36.40	34.51	35.46	44.06	41.41	42.74
MvPLS [6]	41.29	39.34	40.31	42.03	39.40	40.71
GMA [12]	42.26	38.66	40.46	43.96	36.06	40.01
MvDA [2]	39.34	35.04	37.19	41.25	34.58	37.92
MvNDA	43.21	40.81	**42.01**	48.17	42.67	**45.42**
4 views						
MvCCA [6]	40.50	37.91	39.21	45.13	41.66	43.40
MvPLS [6]	41.86	39.74	40.80	41.94	38.84	40.39
GMA [12]	42.26	38.67	40.47	43.30	35.95	39.63
MvDA [2]	41.07	39.21	40.14	41.31	37.16	39.24
MvNDA	43.44	40.63	**42.04**	48.00	42.43	**45.21**

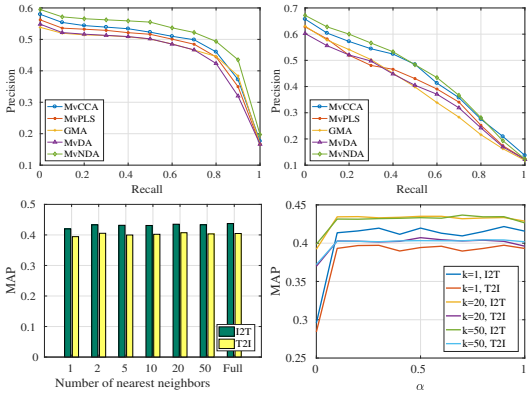


Fig. 3: Clockwise from top left: The precision-recall curve by querying images for text annotations, the retrieval performance of matching text to images, the MAP scores with various α under different fixed numbers of nearest neighors k, (here $k = k_1 = k_2$), and the MAP scores with the different k nearest neighbors and a fixed $\alpha = 0.5$. The legends in the figures in the first row indicate the method producing the PR curve, and we denote querying images for texts by "I2T", and querying texts by images by "T2I" in the figure in the bottom row. k is the number of nearest neighbors.

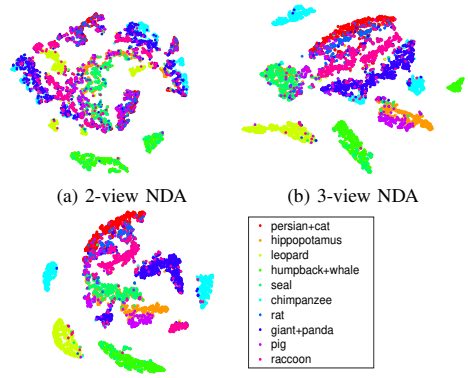
querying every test image and searching for the most relevant texts in the test set, and vice versa. The Mean Average Precision (MAP) is used to evaluate the retrieval performance based on the position of all retrieved images/annotations. We compare the retrieval performance using the features in the subspace of the proposed MvNDA with that of numerous methods in the literature. Both matching images (*view 1*) to text (*view 2*) and text to images are tested. Additional views are projected to the latent space to show more results. In Table I, we see MvNDA outperforms the previous methods in all scenarios using different numbers of views. The further results are confirmed by the Precision-Recall curves in Fig. 3, which shows the retrieval results by the proposed MvNDA are among the leading group in both querying images for text and using text to seek relevant images. We also analyze the effects of different numbers of nearest neighbors and the weight factors α in Fig. 3. It shows the consistent retrieval performance with the different values of k or α, while only using the view-specific discrimination ($\alpha = 0$) degrades the MAP score. We also show the word embedding in its original feature space and the projected latent space in Fig. 1.

B. Animal with Attributes (AwA)

We also demonstrate the effectiveness of multi-view embeddings in tackling the domain shift problem for zero-shot recognition [24]. The Animal with Attribute (AwA) dataset has 50 animal classes with $30,475$ images, and 85 class-level attributes. We follow the experimental protocol in [6] by splitting 40 classes ($24,295$ images) to train the recognition model while the other 10 classes with $6,180$ images for testing

TABLE II: Recognition accuracy (%) on the AwA dataset

	Linear methods			Approxmiate kernel methods		
Method	2 views	3 views	4 views	2 views	3 views	4 views
MvCCA [6]	55.86	75.88	82.01	43.93	47.33	49.51
MvPLS [6]	**58.52**	73.59	77.09	45.37	**47.50**	**52.10**
MvDA [2]	49.95	68.55	70.00	36.65	42.73	42.72
GMA [12]	52.12	73.49	78.46	42.42	44.81	46.84
TMV-HLP [24]	-	73.50	80.50	-	-	-
MvNDA	56.16	**77.16**	**82.78**	**48.78**	46.74	47.56



(a) 2-view NDA (b) 3-view NDA

(c) 4-view NDA

Fig. 4: t-SNE Embedding of Latent Feature Representation: We visualize the embeddings from different numbers of views using the proposed method.

the zero-shot recognition. Each animal class contains more than one positive attribute, and the attributes are shared across classes which enables zero-shot recognition. The detailed class labels and attributes are provided in [25]. Besides the visual features (*view 1,4*) and the class label encoding (*view 3*) generated in the same way as the Wikipedia dataset, a new attribute encoding is added as *view 2* by mapping the visual feature to the attribute probabilities of the animal classes [25].

Table II shows the quantitative results in zero-shot recognition. α, k_1, k_2 are determined based on the grid search using the held-out set. By integrating all available views, we see that recognition accuracy improves with more input views. Due to the size of the training set, we adopt the Nystriöm method for the approximate kernel mapping [10]. MvNDA produces the leading results in all linear cases. We can also observe that the performance of nonlinear methods is inferior compared to the linear ones, which can be explained by the high-dimensionality of the input representations and the use of approximate kernel-based learning. We also graphically show in Fig. 4 that with more available views, the embedded features are grouped into the correct animal classes using the proposed method.

IV. Conclusion

We proposed a novel multi-view nonparametric discriminant analysis technique for the problem of cross-modal image retrieval and recognition. This method has several advantages in exploiting the view difference and class boundary structure information, providing more available projection directions, and achieving better class discrimination in different tasks on both Wikipedia and AwA dataset.

REFERENCES

[1] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.

[2] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[3] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[4] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[5] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 2010, pp. 1–4.

[6] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, 2017, doi: 10.1109/TCYB.2017.2742705.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[8] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, Sept 2013.

[9] ——, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

[10] A. Iosifidis and M. Gabbouj, "Nyström-based approximate kernel subspace learning," *Pattern Recognition*, vol. 57, pp. 190–197, 2016.

[11] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[12] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[14] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, no. 6, pp. 671–678, 1983.

[15] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 31, no. 4, pp. 755–761, 2009.

[16] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[17] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 27, no. 3, pp. 328–340, 2005.

[18] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proceedings of Annual Conference of Computational Learning Theory,*. Springer, Heidelberg, Germany, 2001, pp. 416–426.

[20] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[22] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.

[24] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.

[25] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.

Publication III

G. Cao, M. A. Waris, A. Iosifidis and M. Gabbouj, “Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval,” 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, 2016, pp. 1-6. doi: 10.1109/IPTA.2016.7821032

Multi-modal Subspace Learning with Dropout regularization for Cross-modal Recognition and Retrieval

Guanqun Cao, Muhammad Adeel Waris, Alexandros Iosifidis, Moncef Gabbouj
Dept. of Signal Processing, Tampere University of Technology, Finland
E-mail: {guanqun.cao, muhammad.waris, alexandros.iosifidis, moncef.gabbouj}@tut.fi

***Abstract*—There has been a surge of efforts in cross-modal recognition and retrieval in recent multimedia research. Towards this goal, we investigate a multi-modal subspace learning algorithm together with the Dropout regularizer. Inspired by the regularization for neural networks, we propose to aritificially remove the effect of certain amount of feature bins using the probabilistic approach to prevent the linear subspace learning from over-fitting. The novel regularizer is well integrated into the multi-modal learning algorithm which maximizes the between-class scatter while minimizing the within-class scatter in the projected latent space. The new objective function can be solved efficiently as the generalized eigenvalue problem. Experimental results have shown that superior performance can be obtained in both face-sketch recognition and cross-modal retrieval applications.**

Keywords-subspace learning; face-sketch recognition; cross-modal retrieval;

I. Introduction

Multimedia content is diverse. It has been witnessed recently that, a huge amount of images, videos and annotated texts are generated on a daily basis. There is an increasing need to match the contents sharing the same meaning or purpose across different modalities. One example is identifying criminals based on the forensic sketches. This problem is intrinsically difficult as matching between the criminal faces and the sketches is a challenging issue. Another problem is retrieving the best matching image from a large image database using the query text, or finding the best textual description of an image. This so-call cross-modal image retrieval has also attracted much attention in multimedia research.

Subspace learning method is a type of algorithms to build the relationship between modalities. It aims to project different input features to a common latent space by fitting an optimization criterion. Among these subspace learning methods, Canonical Correlation Analysis (CCA) [1] becomes widely used in establishing pairwise relations for numerous applications [2], [3]. The idea behind CCA is to project different modalities into a latent space, where the inter-modality correlation is maximized and the within-modality correlation is minimized. Another popular technique is Partial Least Square (PLS) regressions [4], which maximize the covariance between modalities only. Recent attempts have been focused on exploiting the class discrimination. Multi-view Discriminant Analysis [5] is a technique which extends Linear Discriminant Analysis (MvDA) [6] to project more than two input modalities into the latent space. On top of MvDA, Joint Feature Selection and Subspace learning (JFSSL) [7] is proposed for multi-modal learning with class discriminant in the regularization term.

There are certain limitations in the existing subspace learning techniques. Traditional methods such as CCA and PLS accept only pairwise information, and it has been shown to provide performance gain from more modalities [8]. In MvDA, the between-class scatter is maximized regardless of the difference between inter-modality and intra-modality covariances, while the within-class scatter is minimized in the mean time. JFSSL has taken the cross-modality relation into account by building a multi-modal graph regularization term. However, the modality discrimination is only considered in the regularization step. The class and modality information is not well integrated and therefore the potential of joint optimization cannot be sufficiently exploited. In contrast, the Dropout regularizer inpired in neural networks [9], [10] can be a simple yet powerful alternative to penalize the modality difference and avoid over-fitting for linear subspace learning.

In this paper, we propose a Dropout regularized multi-modal subspace learning algorithm based on [11] for face-sketch recognition and cross-modal retrieval. The novel method aims to reduce the modality discrepancy when projecting the multiple features into a common space, which makes the learning algorithm robust in case of over-fitting. We introduce the Dropout regularizer for linear subspace learning which randomly removes the effect of certain number of feature vector bins. This artificially corrupting term is well combined with the inter-modality and intra-modality scatters to become a unified objective function. The proposed formulation integrates the cross-modal scatters and cross-class covariances as a whole, and can be solved efficiently as a generalized eigenvalue problem. We conduct a series of experiments to show the proposed learning algorithm consistently achieves superior results in both recognition and retrieval applications.

The rest of the paper is organized as follows. In Section II, we describe explicitly the joint multi-modal subspace learning with Dropout. Then, in Section III, we present the

comparative results in face-sketch recognition and cross-modal image retrieval on two popular multimedia datasets. Finally, Section IV concludes the paper.

II. Multi-modal subspace learning with Dropout regularization

In this section, a novel multi-modal subspace learning algorithm with Dropout regularization is presented. We firstly introduce the regularization term. Then the multi-modal subspace learing is described. Finally, we show the integration of the regularization into the multi-modal objective function, and its solution as a generalized eigenvalue problem.

A. Dropout regularization for linear subspace learning

We define the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, $\mathbf{x}_i \in \mathbb{R}^D$, where N is the number of samples and D is the feature dimension. We also define $\mathbf{X}_v \in \mathbb{R}^{D_v \times N}, v = 1, \ldots, V$ for the feature vectors of the vth view. Note that the dimensionality of the various feature spaces D_v may vary across the views. $\mathbf{W}_v \in \mathbb{R}^{D_v \times d}, v = 1, \ldots, V$ is the weight matrix for each view, d is the number of dimensions in the latent space. After the linear projection, we can obtain the latent feature vectors from each view as

$$\mathbf{Y}_v = \mathbf{W}_v^\top \mathbf{X}_v. \tag{1}$$

The idea of Dropout regularization is to create a binary mask $\mathbf{m}_{i,t}$ to remove certain number of values from the original feature vector at each epoch. It originates from dropping out neurons and their connections while training the neural networks. The elements in $\mathbf{m}_{i,t}$ equals to 1 with a probability value p following a Bernoulli distribution, and equals to 0 with $(1-p)$ probability. Then the feature vector after applying Dropout is

$$\mathbf{x}_{i,t} = \mathbf{m}_{i,t} \circ \mathbf{x}_i, \tag{2}$$

where $\circ$ denotes the operation for the Hadamard (element-wise) product. We also express the feature vector for shrinkage as $\widetilde{\mathbf{x}}_{i,t} = \mathbf{x}_i - \mathbf{x}_{i,t}$, then it is easy to derive the regularized feature vector in the latent space as

$$\mathbf{W}^\top (\mathbf{x}_{v,i} - \mathbf{x}_{v,i,t}) = \mathbf{W}^\top \widetilde{\mathbf{x}}_{v,i,t} = \mathbf{0}. \tag{3}$$

The regularization term is shown as

$$R(\mathbf{W}) = \frac{1}{2\,N_T} \sum_{v=1}^{V} \sum_{i=1}^{N} \sum_{t=1}^{N_T} \|\mathbf{W}_v^\top \mathbf{x}_{v,i} - \mathbf{W}_v^\top \mathbf{x}_{v,i,t}\|_F^2 \tag{4}$$

$$= \frac{1}{2\,N_T} \sum_{v=1}^{V} \sum_{t=1}^{N_T} \|\mathbf{W}_v^\top \widetilde{\mathbf{X}}_{v,t}\|_F^2 \tag{5}$$

We consider the case when the number of epoches N_T goes to infinity. Then from the weak law of large numbers, we know that $R(\mathbf{W})$ will converge to its expectation

$$R(\mathbf{W}) = \frac{1}{2\,N_T} \sum_{v=1}^{V} \sum_{t=1}^{N_T} E\left(\mathbf{W}_v^\top \widetilde{\mathbf{X}}_{v,t} \widetilde{\mathbf{X}}_{v,t}^\top \mathbf{W}_v\right) \tag{6}$$

$$= \frac{1}{2\,N_T} \sum_{v=1}^{V} \sum_{t=1}^{N_T} \mathbf{W}_v^\top \left(\widetilde{\mathbf{X}}_{v,t} \widetilde{\mathbf{X}}_{v,t}^\top \circ \mathbf{P}\right) \mathbf{W}_v, \tag{7}$$

where $\mathbf{P} = \left[(\mathbf{p}\mathbf{p}^\top) \circ (\mathbf{1}\mathbf{1}^\top - \mathbf{I})\right] + \left[(\mathbf{p}\mathbf{I}^\top) \circ \mathbf{I}\right]$, and $\mathbf{p} = [(1-p), \ldots, (1-p)]^\top \in \mathbb{R}^N$ is a vector whose elements shows the probability that $\mathbf{x}_i = 0$. We also define that $\mathbf{1} \in \mathbb{R}^{N \times N}$ as a vector of ones, and $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix.

B. Multi-modal subspace learning

Here we propose the expression of objective function for multi-view subspace learning as

$$\begin{aligned} &\textbf{Maximize}: && \frac{\mathrm{Tr}(\mathbf{S}_B)}{\mathrm{Tr}(\mathbf{S}_W + \alpha R(\mathbf{W}))} \\ &\Longrightarrow && \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{P} \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{Q} \mathbf{W} + \alpha R(\mathbf{W}))}. \\ &\textbf{Subject to}: && \mathbf{W}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{8}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the matrices describing the between-class and within-class scatters, respectively. $\mathbf{P}$ and $\mathbf{Q}$ are the inter-view and intra-view covariance matrices. α is the parameter adjusting the importance of regularization.

We exploit the Dropout regularization of the objective function in the case of linear projection. $\mathbf{S}_{vij}$ is a similarity weight matrix which encodes the intra-view properties to be minimized, and $\mathbf{S}'_{vij}$ is a penalty weight expressing the inter-view properties to be maximized. The weight matrix $\mathbf{W}$ is thereby obtained from

$$\begin{aligned} &\textbf{Maximize}: && \frac{\sum_{v=0}^{V} \sum_{i=0}^{N} \sum_{j=0}^{N} \mathbf{S}'_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2}{\sum_{v=0}^{V} \sum_{i=0}^{N} \sum_{j=0}^{N} \mathbf{S}_{vij} \|\mathbf{W}_v^\top \mathbf{X}_{vi} - \mathbf{W}_v^\top \mathbf{X}_{vj}\|^2 + \alpha R(\mathbf{W})} \\ &\Longrightarrow && \frac{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L}_B \mathbf{X}^\top \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^\top \mathbf{X} \mathbf{L}_W \mathbf{X}^\top \mathbf{W} + \alpha R(\mathbf{W}))}, \\ &\textbf{Subject to}: && \mathbf{W}^\top \mathbf{W} = \mathbf{I} \end{aligned}$$

In the above, we define the diagnal matrix of each view pair as $\mathbf{D}_{uv}$ whose i-th element is $[\mathbf{D}_{uv}]_{ii} = \sum_j [\mathbf{S}_{uv}]_{ij}$, and the total graph Laplacian matrix as $\mathbf{L}_B = \mathbf{D} - \mathbf{S}$. Similarly, we have $\mathbf{D}', \mathbf{S}', \mathbf{L}_W$ in the penalty graph.

We propose the between-class scatter matrix which maximizes the distance between different class centers of different views as

$$\mathbf{S}_B = \sum_{i=1}^{V} \sum_{j=1}^{V} \sum_{p=1}^{C} \sum_{\substack{q=1 \\ p \neq q}}^{C} (\boldsymbol{m}_p^i - \boldsymbol{m}_q^i)(\boldsymbol{m}_p^j - \boldsymbol{m}_q^j)^\top$$

$$= \sum_{i=1}^{V} \sum_{j=1}^{V} \sum_{p=1}^{C} \sum_{q=1}^{C} \mathbf{W}_i^\top \mathbf{X}_i \mathbf{L}_B \mathbf{X}_j^\top \mathbf{W}_j, \tag{9}$$

and the Laplacian matrix is

$$\mathbf{L}_B = 2\,(\mathbf{e}_p\,\mathbf{e}_p^\top - \mathbf{e}_p\,\mathbf{e}_q^\top), \tag{10}$$

where e_p, e_q are the class vectors, and both are defined as $\mathbf{e}^p \in \mathbb{R}^N$ with $e_p(i) = 1$, if $p_i = p$, and $e_p(i) = 0$, otherwise. We define the formulation of within-class Laplacian matrix as

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{V}\sum_{c=1}^{C} \mathbf{W}_i^\top \mathbf{X}_i \left(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top\right)\mathbf{X}_i^\top \mathbf{W}_i \\ &= \sum_{i=1}^{V}\sum_{i=1}^{V}\sum_{c=1}^{C} \mathbf{W}_i^\top \mathbf{Q}_{ii}\,\mathbf{W}_i, \end{aligned} \tag{11}$$

The above equation has the form of the Rayleigh quotient. Therefore, all subspace learning methods that maximize the criterion are reduced to a generalized eigenvalue problem:

$$\begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1V} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{V1} & \mathbf{P}_{V2} & \cdots & \mathbf{P}_{VV} \end{bmatrix} \mathbf{W} = \boldsymbol{\rho} \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{VV} \end{bmatrix} \mathbf{W}, \tag{12}$$

and the solution is given below:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_V \end{pmatrix} \text{ and } \boldsymbol{\rho} = \sum_{i=1}^{d} \lambda_i \tag{13}$$

are the generalized eigenvector and the sum of the top d generalized eigenvalues λ_i respectively. $\mathbf{W}$ contains the projection matrices of all views, and $\boldsymbol{\rho}$ is the value of Raleigh quotient. We take the top d dimensions of eigenvectors into the projection matrix, and their corresponding eigenvalues for $\boldsymbol{\rho}$ as the solution. We address the Rayleigh quotient as the uniform objective function, reaching out to all subspace learning methods in the paper.

III. Experimental Results

We perform several experiments to demonstrate the effectiveness of the proposed method in multi-modal face-sketch recognition and cross-modal image retrieval tasks. Firstly, we describe the procedure of conducting each experiment, the features that we employed, and the evaluation metrics. Then, we show the comparative results of matching between faces and sketches by algorithms including CCA [12], PLS [13], MvDA [5] and the propose method. We continue to evaluate the performance of multi-modal learning in cross-modal retrieval. In each section, we present figures and analysis of the results and discuss the effect of using the proposed method on different datasets.

A. Experimental Setup

In the Face-Sketch recognition experiment, we find the best matching face to each sketch from the test database and vice versa. Facial feature and sketch features are extracted in their own domain. Specifically, all images are aligned based on their fiducial points and cropped to 80×64 pixels. We use the pre-trained CNN models produced by deep residual net [14] to generate the facial feature. The output from the '*data*' layer is used as the feature vectors. The sketches are represented by Histogram of oritented Gradients (HoG) feature [15]. Principal Component Analysis (PCA) is applied to reduce the dimensionality of both face and sketch features. Empirically, we the set number of dimensionality to 500 to show the best recognition accuracies. Thereafter, we train a model by projecting both features into a common space, and compute the matching between new faces and sketched in the latent space.

CUHK Face Sketch FERET (CUFSF) [16], [2] is used to evaluated the Face-Sketch recognition. The dataset is consist of $1,194$ subjects with lighting variations from FERET dataset [17]. Each subject is represented by a pair of face and sketch, and each sketch is drawn manually with shape exaggeration according to its face image. We use the first 700 subjects for training and the rest for testing. The evaluation of matching different modalities is presented quantitatively using rank-1 recognition rate.

In cross-modal retrieval, we perform both text query for image retrieval and image query for text retrieval. Furthermore, the experiments are extended to 2-view, 3-view and 4-view cases and the retrieval performance are compared between numerous algorithms. We first extract the query features in their own domains. We learn a joint model from the latent space using the image features, class label features and sentence features. Query features are then projected to this latent space. We also undertake a semantic matching step [3] which trains a logistic regressor over the embedded features from all of the ground truth samples mapping the projected features of both queries or to-be-retrieved images/texts towards the class labels. The feature vectors generated from the ground truth class labels is essentially the class vectors, whose number of dimension is the number of classes. We finally use the class probabilities based on the logistic regressor outputs to perform the cross-modal retrieval.

We use the following feature representations in the experiments:

- **Image feature by CNN models**: We again employ the off-the-shelf CNN models to extract the visual features. We firstly rescale the size of the images to 224×224, and set the mini-batch size to 50. We use the output from the $fc8$ layer in VGGNet with 16 weight layers [18] (denoted as *VGG-16* in latter sections) and the $loss3/classifier$ layer from GoogleNet [19]. Both models produce 1000-dimension feature vectors.
- **Class label encoding**: Since each image corresponds to one class label on the Wikipedia dataset, we exploit the textual feature from the category names mapped from the image features. Specifically, we firstly train a 100-

dimension skip-gram model [20] on the entire English Wikipedia articles composed of 2.9 billion words. Then we can extract a separate set of word vectors from class labels of our datasets. In order to correlate the word and images, we train a ridge regressor with 10-fold cross-validation to map the *VGG-16* image features to each dimension of the word vectors respectively. The regressor outputs are used as the class label features.

- **Sentence encoding**: A vital feature of cross-modal retreival system is that we make use of textual features directly. We can find a paragraph of text describing each image on the Wikipedia dataset. We generated the sentence vectors from the paragraphs by the pre-trained skip-thoughts model [21]. The model was trained over the MovieBook and BookCorpus dataset [22]. The resultant feature extracted by the Skip-thought model 4800 dimensions.

Wikipedia is a cross-modal dataset collected from the "Wikipedia featured articles" [3]. The dataset is organized in 10 categories and consists of $2,866$ documents. Each document is a short paragraph with a median text length of 200 words, and is associated with a single image. We follow the train/test split in [3] who use $2,173$ training and 693 test pairs of images and documents. PCA is also applied in before the subspace learning. and we set the number of dimension to 100.

We present the retrieval results graphically using 11-point interpolated precision-recall (PR) curves. The Mean Average Precision (MAP) score, which is the average precision at the ranks where recall changes, can be computed based on the Precision Recall curves. The Average Precision (AP) measures the relevance between a query and retrieved items [23], and the MAP score calculates the mean AP by quering all items in the test set.

B. Results on Face-Sketch Recognition

TABLE I: Recognition Rate (%) on the CUFSF Dataset

Method	Face-Sketch	Sketch-Face	Avg.
CCA	48.79	52.83	50.81
PLS	31.38	31.38	31.38
MvDA	45.55	49.60	47.58
LDA	47.17	51.62	49.40
LDA-Dropout	61.13	64.98	**63.06**

Table I shows the accuracy for both face-sketch and sketch-face recognition by CCA, PLS, MvDA, and the proposed LDA without and with the Dropout regularization. It can be seen that the proposed algorithm in the last row outperforms the relative methods by a large margin. It exhibits the property of adoption of supervised information and robustness against over-fitting. The matching between different modalities benefits from the discrimination and regularization. Moreover, we present a graph showing the influence of the probability p on the recognition performance at different levels of regularization importance (α). We observe that the recognition rate is generally consistent to different Dropout probabilities, and always better than the one without the regularization, i.e. $p = 1$.

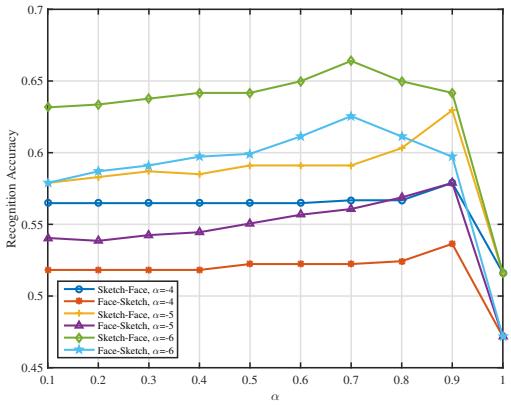


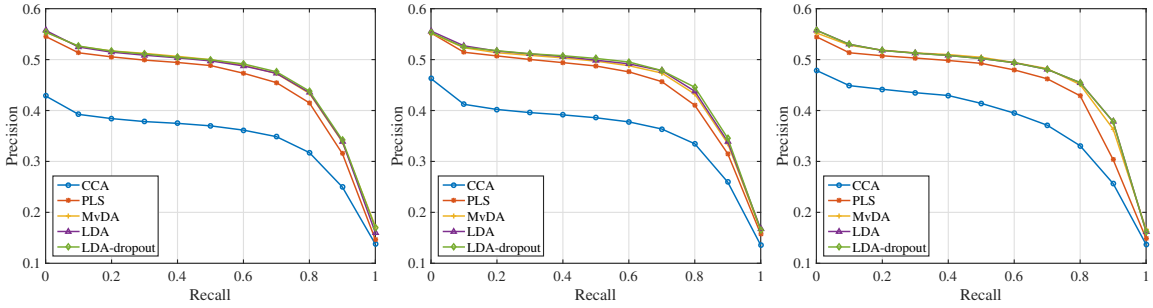
Fig. 1: Face-Sketch Recognition Rate for different probability p.

C. Results on Cross-modal Retrieval

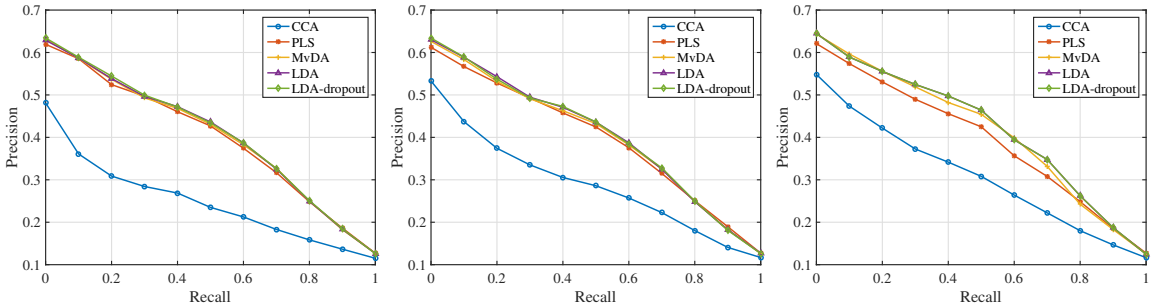
We use the *VGG-16* features and sentence features for two views, and augment class label and GoogleNet encodings as the additional modalities. The quantitative results are shown in Table II. It can be seen that a better retrieval is obtained when enriching the latent feature with more modalities. We also observe that the discriminant information improve the performance, and both text-image and image-text retrieval progress with more modalities. Moreover, the proposed methods with Dropout is the best algorithm in all categories of using different number of modalities. We present more detailed results in the form of PR curves in Fig. 2. It can be seen that the proposed method consistently outperforms the other methods across all views, while most methods are comparable except CCA.

IV. Conclusion

We have proposed a novel Dropout regularized multi-modal subspace learning algorithm. The regularizer artificially generates zero feature values to penalize the view difference and avoid over-fitting during linear projections. The expression for regularization is a natural extension in neural networks, and is well integrated in the unified objective function .The joint optimization formulation can maximize the inter-modality scatters and minimize the intra-modality scatters. Meanwhile, the between-class covariance are maximized while within-class covariances are minimized. The formulation can be solved efficiently as a generalized eigenvalue problem. We conducted several experiments in Face-Sketch recognition and Cross-modal retrieval. In both applications, the results have shown that the proposed algorithm achieves consistently superior results against state-of-the-art methods.



(a) PR curves for image queries with 2 views (b) PR curves for image queries with 3 views (c) PR curves for image queries with 4 views



(d) PR curves for text queries with 2 views (e) PR curves for text queries with 3 views (f) PR curves for text queries with 4 views

Fig. 2: PR curves across different number of views on the Wikipedia dataset for the Image-to-Text retrieval and the Text-to-Image retrieval.

TABLE II: MAP Score (%) on the Wikipedia Dataset

Method	img. query	txt. query	Avg.
2 views			
CCA	29.93	22.62	26.28
PLS	40.51	38.05	39.28
MvDA	42.12	38.65	40.38
LDA	41.96	38.80	40.38
LDA-Dropout	42.32	38.93	**40.63**
3 views			
CCA	31.99	27.00	29.50
PLS	40.76	37.85	39.31
MvDA	41.94	38.48	40.21
LDA	42.28	38.81	40.55
LDA-Dropout	42.40	38.79	**40.60**
4 views			
CCA	33.55	28.79	31.17
PLS	40.72	37.69	39.21
MvDA	43.20	39.74	41.47
LDA	43.46	39.94	41.70
LDA-Dropout	43.47	39.94	**41.71**

REFERENCES

[1] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[2] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 513–520.

[3] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 521–535, 2014.

[4] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 593–600.

[5] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[7] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. PP, no. 99, pp. 1–1, 2015.

[8] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 37, no. 11, pp. 2332–2345, Nov 2015.

[9] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Dropelm: Fast neural network regularization with dropout and dropconnect," *Neurocomputing*, vol. 162, pp. 57–66, 2015.

[11] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-

view embedding for visual recognition and cross-modal retrieval," *arXiv preprint arXiv:1605.09696*, 2016.

[12] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing (TIP)*, vol. 11, no. 3, pp. 293–305, 2002.

[13] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*. Springer, 2006, pp. 34–51.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2015.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, vol. 1, 2005, pp. 886–893.

[16] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, Nov 2009.

[17] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2015, pp. 1–9.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.

[21] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3276–3284.

[22] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," *arXiv preprint arXiv:1506.06724*, 2015.

[23] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to Information Retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1, ch. 8. Evaluation in information retrieval, pp. 188–210.

Publication IV

G. Cao, A. Iosifidis, M. Gabbouj, V. Raghavan, R. Gottumukkala, Deep Multi-view Learning to Rank, submitted to IEEE Trans. on Knowledge and Data Engineering. arXiv:1801.10402

Deep Multi-view Learning to Rank

Guanqun Cao, Alexandros Iosifidis, Moncef Gabbouj, Vijay Raghavan, Raju Gottumukkala

Abstract—We study the problem of learning to rank from multiple sources. Though multi-view learning and learning to rank have been studied extensively leading to a wide range of applications, multi-view learning to rank as a synergy of both topics has received little attention. The aim of the paper is to propose a composite ranking method while keeping a close correlation with the individual rankings simultaneously. We propose a multi-objective solution to ranking by capturing the information of the feature mapping from within each view as well as across views using autoencoder-like networks. Moreover, a novel end-to-end solution is introduced to enhance the joint ranking with minimum view-specific ranking loss, so that we can achieve the maximum global view agreements within a single optimization process. The proposed method is validated on a wide variety of ranking problems, including university ranking, multi-view lingual text ranking and image data ranking, providing superior results.

Index Terms—Learning to rank, multi-view data analysis, ranking

1 Introduction

Learning to rank is an important research topic in information retrieval and data mining, which aims to learn a ranking model to produce a query-specfic ranking list. The ranking model establishes a relationship between each pair of data samples by combining the corresponding features in an optimal way [1]. A score is then assigned to each pair to evaluate its relevance forming a global ranking list across all pairs. The success of learning to rank solutions has brought a wide spectrum of applications, including online advertising [2], natural language processing [3] and multimedia retrieval [4].

Learning appropriate data representation and a suitable scoring function are two vital steps in the ranking problem. Traditionally, a feature mapping models the data distribution in a latent space to match the relevance relationship, while the scoring function is used to quantify the relevance measure [1]; however, the ranking problem in the real world emerges from multiple facets and data patterns are mined from diverse domains. For example, universities are positioned differently based on numerous factors and weights used for quality evaluation by different ranking agencies. Therefore, a global agreement across sources and domains should be achieved while still maintaining a high ranking performance.

Multi-view learning has received a wide attention with a special focus on subspace learning [5], [6] and co-training [7], and few attempts have been made in ranking problems [8]. It introduces a new paradigm to jointly model and combine information encoded in multiple views to enhance the learning performance. Specifically, subspace learning finds a common space from different input modalities using an optimization criterion. Canonical Correlation Analysis (CCA) [9], [10] is one of the prevailing unsupervised method used to measure a cross-view correlation. By contrast, Multi-view Discriminant Analysis (MvDA) [6] is a supervised learning technique seeking the most discriminant features across views by maximizing the between-class scatter while minimizing the within-class scatter in the underlying feature space. Furthermore, a generalized multi-view embedding method [5] was proposed using a graph embedding framework for numerous unsupervised and supervised learning techniques with extension to nonlinear transforms including (approximate) kernel mappings [11], [12] and neural networks [5], [13]. A nonparametric version of [5] was also proposed in [14]. On the other hand, co-training [7] was introduced to maximize the mutual agreement between two distinct views, and can be easily extended to multiple inputs by subsequently training over all pairs of views. A solution to the learning to rank problem was provided by minimizing the pairwise ranking difference using the same co-training mechanism [8].

Although there are several applications that could benefit from multi-view learning to rank approach, the topic has still been insufficiently studied up to date [15]. Ranking of multi-facet objects is generally performed using composite indicators. The usefulness of a composite indicator depends upon the selected functional form and the weights associated with the component facets. Existing solutions for university ranking are an example of using the subjective weights in the method of composite indicators. However, the functional form and its assigned weights are difficult to define. Consequently, there is a high disparity in the evaluation metric between agencies, and the produced ranking lists usually cause dissension in academic institutes. However, one observation is that, the indicators from different agencies may partially overlap and have a high correlation between each other. We present an example in Fig. 1 to show that, several attributes in the THE dataset [16], including teaching, research, student staff ratio and student number are highly correlated with all of the attributes in the ARWU dataset [17]. Therefore, the motivation of this paper is to find a composite ranking by exploiting the correlation between individual rankings.

Earlier success in multi-view subspace learning provides a promising way for composite ranking. Concatenating multiple views into a single input overlooks possible view

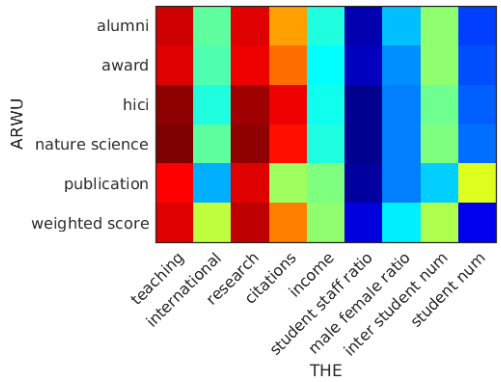


Fig. 1: The correlation matrix between the measurements of Times Higher Education (THE) and Academic Ranking of World Universities (ARWU) rankings. The data is extracted and aligned based on the performance of the common universities in 2015 between the two ranking agencies. The reddish color indicates high correlation, while the matrix elements with low correlation are represented in bluish colors.

discrepancy and does not fully exploit their mutual agreement in ranking. Our goal is to study beyond the direct multi-view subspace learning for ranking. This paper offers a multi-objective solution to ranking by capturing relevant information of feature mapping from within each view as well as across views. We propose a generic framework for multi-view subpsace learning to rank (MvSL2R). It incorporates novel feature embedding methods of both multi-view unsupervised and discriminant autoencoders. Moreover, we propose an end-to-end method to optimize the trade-off between view-specific ranking and a discriminant combination of multi-view ranking. To this end, we can improve cross-view ranking performance while maintaining individual ranking objectives.

Intermediate feature representation in the neural network are exploited in our ranking solutions. Specifically, the first contribution is to provide two closely related methods by adopting an autoencoder-like network. We first train a network to learn view-specific feature mappings, and then maximize their correlation with the intermediate representations using either an unsupervised or discriminant projection to a common latent space. A stochastic optimization method is introduced to fit the correlation criterion. Both the autoencoding sub-network per view with a reconstruction objective and feedforward sub-networks with a joint correlation-based objective are iteratively optimized in the entire network. The projected feature representations in the common subspace are then combined and used to learn for the ranking function.

The second contribution (graphically described in Fig. 2) is an end-to-end multi-view learning to rank solution. A sub-network for each view is trained with its own ranking objective. Then, features from intermediate layers of the sub-networks are combined after a discriminant mapping to a common space, and training towards the global ranking objective. As a result, a network assembly is developed to enhance the joint ranking with mimimum view-specific ranking loss, so that we can achieve the maximum view agreement within a single optimization process.

The rest of the paper is organized as follows. In Section 2, we describe the related work close to our proposed methods. The proposed methods are introduced in Section 3. In Section 4, we present quantitative results to show the effectiveness of the proposed methods. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Learning to rank

Learning to rank aims to optimize the combination of data representation for ranking problems [18]. It has been widely used in a number of applications, including image retrieval and ranking [4], [19], image quality ratings [20], online advertising [2], and text summarization [8]. Solutions to this problem can be decomposed into several key components, including the input feature, the output vector and the scoring function. The framework is developed by training the scoring function from the input feature to the output ranking list, and then, scoring the ranking of new data. Traditional methods also include engineering the feature using the PageRank model [21], for example, to optimally combine them for obtaining the output. Later, research was focused on discriminatively training the scoring function to improve the ranking outputs. The ranking methods can be organized in three categories for the scoring function: the pointwise approach, the pairwise apporach, and the listwise approach.

We consider the pairwise approach in this paper and review the related methods as follows. A preference network is developed in [22] to evaluate the pairwise order between two documents. The network learns the preference function directly to the binary ranking output without using an additional scoring function. RankNet [23] defines the cross-entropy loss and learns a neural network to model the ranking. Assuming the scoring function to be linear [24], the ranking problem can be transformed to a binary classification problem, and therefore, many classifiers are available to be applied for ranking document pairs. RankBoost [25] adopts Adaboost algorithm [26], which iteratively focuses on the classfication errors between each pair of documents, and subsequently, improves the overall output. Ranking SVM [27] applies SVM to perform pairwise classification. GBRank is a ranking method based on Gradient Boost Tree [28]. Semi-supervised multi-view ranking (SmVR) [8] follows the co-training scheme to rank pairs of samples. Moreover, recent efforts focus on using the evaluation metric to guide the gradient with respect to the ranking pair during training. These studies include AdaRank [29], which optimizes the ranking errors rather than the classification error in an adaptive way, and LambdaRank [30]. However, all of these methods above consider the case of single view inputs, while limited work on multi-view learning to rank has been studied [15], [31], [32].

2.1.1 Bipartite ranking

The pairwise approach of the ranking methods serves as the basis of our ranking method, and therefore, reviewed explicitly in this section. Suppose that the training data is organized in query-sample pairs $\{(\mathbf{x}_i^q, \mathbf{y}_i^q)\}$, where $q \in \{1, 2, \ldots, Q\}$, $\mathbf{x}_i^q \in \mathrm{R}^d$ is the d-dimensional feature vector for the pair of query q, the i-th sample, $\mathbf{y}_i^q \in \{0, 1\}$ is the relevance score, and the number of query-specific samples is N_q. We perform the pairwise transformation before the relevance prediction of each query-sample pair, so that only the samples that belong to the same query are evaluated [24].

The modeled probability between each pair in this paper is defined as

$$\mathbf{p}_i^q(\phi) = \frac{1}{1 + \exp(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_q))},$$

where $\phi : \mathbf{x} \to \mathbb{R}$ is the linear scoring function as $\phi(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$, which maps the input feature vectors to the scores. Due to its linearity, we can transform the feature vectors and relevance score into $(\mathbf{x}'_k, \mathbf{y}'_k) = (\mathbf{x}_q - \mathbf{x}_i, \mathbf{y}_i^q)$. In case of the ordered list ($\mathbf{r}$) as the raw input, each data sample $\mathbf{x}_i$ paired with its query $\mathbf{x}_q$ is investigated, and their raw orders $(\mathbf{r}_i, \mathbf{r}_q)$ are transformed as $\mathbf{y}_i^q = 1$, if $\mathbf{r}_i < \mathbf{r}_q$; $\mathbf{y}_i^q = 0$, else if $\mathbf{r}_i > \mathbf{r}_q$. In pairwise ranking, the relevance $\mathbf{y}_i^q = 1$, if the query and sample are relevant, and $\mathbf{y}_i^q = 0$, otherwise.

The feature difference $(\mathbf{x}'_k, \mathbf{y}'_k)$ becomes the new feature vector as the input data for nonlinear transforms and subspace learning. Therefore, the probability can be rewritten as

$$\mathbf{p}_k(\phi) = \frac{1}{1 + \exp(-\phi(\mathbf{x}'_k))} = \frac{1}{1 + \exp(-\mathbf{a}^\top \mathbf{x}'_k)}. \quad (1)$$

The objective to make the right order of ranking can then be formulated as the cross entropy loss such that,

$$\begin{aligned} \ell_{\text{Rank}} &= \arg\min \sum_{q=1}^{Q} \sum_{i=1}^{N_q} \left(\mathbf{y}_i^q \log \mathbf{p}_i^q) + (1 - \mathbf{y}_i^q) \log \mathbf{p}_i^q)\right) \\ &= \arg\min \sum_{k=1}^{K} \left(\mathbf{y}'_k \log \mathbf{p}_k) + (1 - \mathbf{y}'_k) \log \mathbf{p}_k)\right), \end{aligned} \quad (2)$$

which is proved in [23] that it is an upper bound of the pairwise 0-1 loss function and optimized using gradient descent. The logistic regression or softmax function in neural networks can be used to learn the scoring function.

2.2 Multi-view deep learning

Multi-view learning considers enhancing the feature discriminability by taking inputs from diverse sources. One important approach to follow is subspace learning, which is traced back to CCA [33], [34] between two input domains, and its multi-view extension, which has been studied in [35], [36], [37]. This approach can also be generalized using a higher-order correlation [37]. The main idea behind these techniques is to project the data representations in the two domains to a common subspace optimizing their mutual correlation. Subspace learning with supervision has also been extensively studied. Multi-view Discriminant Analysis [6] performs the dimensionality reducation of features from multiple views exploiting the class information. Recently, these methods were generalized in the same framework [5], [38], which accommodates multiple views, supervision and nonlinearity. Co-training [7] first trains two seperate regressors and then, iteratively maximizes their agreements.

Deep learning, which exploits the nonlinear transform of the raw feature space, has also been studied in the multi-view scenario. The multi-modal deep autoencoder [39] was proposed by taking nonlinear representations of a pair of views to learn their common characteristics. Deep CCA [13] is another *two-view* method which maximizes the pairwise correlation using neural networks. Thereafter, a two-view correlated autoencoder was developed [40], [41] with objectives to correlate the view pairs but also reconstruct the individual view in the same network. By contrast, we propose a generic framework which is extensible for multiple views and both unsupervised and discriminant autoencoders for ranking. The previous work on discriminant autoencoders introduced an additional regularization term [42], while our method embeds the class discrimination in the Laplacian matrix, and is extensible for multiple views.

Multi-view Deep Network [43] was also proposed as an extension of MvDA [6]. It optimizes the ratio trace of the graph embedding [44] to avoid the complexity of solutions without a closed form [45]. In this paper, however, we show that the trace ratio optimization can be solved efficiently in the updates of the multi-view networks. Deep Multi-view Canonical Correlation Analysis (DMvCCA) and Deep Multi-view Modular Discriminant Analysis (DMvMDA) [5] are closely related to our work, and hence, they are described in the following sections.

2.2.1 Deep Multi-view Canonical Correlation Analysis (DMvCCA)

The idea behind DMvCCA [5] is to find a common subspace using a set of linear transforms $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_v]^\top$ to project nonlinearly mapped input samples $\mathbf{Z}_v$ from the vth view where the correlation is maximized. Specifically, it aims to maximize

$$\mathcal{J}_{\text{DMvCCA}} = \underset{\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L} \mathbf{Z}_i^\top \mathbf{W}_i = \mathbf{I}}{\arg\max} \operatorname{Tr}\left(\sum_{i=1}^{V} \sum_{\substack{j=1 \\ j \neq i}}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L} \mathbf{Z}_j^\top \mathbf{W}_j\right) \quad (3)$$

where the matrix $\mathbf{L} = \mathbf{I} - \frac{1}{N}\mathbf{e}\mathbf{e}^\top$ centralizes the input data matrix of each view v, and $\mathbf{e}$ is a vector of ones, . By defining the cross-view covariance matrix between views i and j as $\mathbf{\Sigma}_{ij} = \frac{1}{N}\tilde{\mathbf{Z}}_i \tilde{\mathbf{Z}}_j$, where $\tilde{\mathbf{Z}}_v, v = 1, \ldots, V$, is the centered view, the data projection matrix $\mathbf{W}$, which has the column vector of $\mathbf{W}_v$ in the vth view, can be obtained by solving the generalized eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1V} \\ \mathbf{\Sigma}_{21} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{V1} & \mathbf{\Sigma}_{V2} & \cdots & \mathbf{0} \end{bmatrix} \mathbf{W} = \lambda \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Sigma}_{VV} \end{bmatrix} \mathbf{W}. \quad (4)$$

It shows that the solution to this problem is derived with the maximal inter-view covariances and the minimal intra-view

covariances.

2.2.2 Deep Multi-view Modular Discriminant Analysis (DMvMDA)

DMvMDA [5] is the neural network-based multi-view solution of LDA which maximizes the ratio of the determinant of the between-class scatter matrix of all view pairs to that of the within-class scatter matrix. Mathematically, it is written as the projection matrix of the DMvMDA and is derived by optimizing function

$$\mathcal{J}_{\text{DMvMDA}} = \underset{\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_W \mathbf{Z}_i^\top \mathbf{W}_i = \mathbf{I}}{\arg\max} \text{Tr}\,(\sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_B \mathbf{Z}_j^\top \mathbf{W}_j), \tag{5}$$

where the between-class Laplacian matrix is

$$\mathbf{L}_B = 2\sum_{p=1}^{C}\sum_{q=1}^{C}(\frac{1}{N_p^2}\mathbf{e}_p\,\mathbf{e}_p^\top - \frac{1}{N_p N_q}\mathbf{e}_p\,\mathbf{e}_q^\top).$$

and the within-class Laplacian matrix is

$$\mathbf{L}_W = \mathbf{I} - \sum_{c=1}^{C}\frac{1}{N_c}\mathbf{e}_c\mathbf{e}_c{}^\top.$$

3 Model Formulation

We first introduce multi-view subspace learning to rank (MvSL2R). It is followed by the formulations of MvCCAE and MvMDAE. Finally, the end-to-end ranking method is presented.

3.1 Multi-view Subspace Learning to Rank (MvSL2R)

Multi-view subspace learning to rank is formulated based on the fact that the projected feature in the common subspace can be used to train a scoring function for ranking. We generate the training data from the intersection of ranking samples between views to have the same samples but various representations from different view origins. The overall ranking agreement is made by calculating the average voting from the intersected ranking orders as

$$\bar{\mathbf{r}} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{r}_v. \tag{6}$$

By performing the pairwise transform in section 2.1.1 over the ranking data, we have the input $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ of V views and the cross-view relevance scores $\bar{\mathbf{y}}$ obtained from the average ranking orders $\bar{\mathbf{r}}$. The proposed ranking method consists of feature mapping into a common subspace, training a logistic regressor as the scoring function, and predicting the relevance of new sample pairs using the probability function

$$\mathbf{p}_v(\mathbf{X}_v) = \frac{1}{1+\exp(-\mathbf{a}^\top \mathbf{W}_v^\top \mathcal{F}_v(\mathbf{X}_v))}, \tag{7}$$

where $\mathbf{W}_v$ is the data projection matrix of the vth view, and $\mathbf{a}$ is the weight from the logistic regressor described in (1). We summarize these steps in the algorithm below.

Algorithm 1: Multi-view Subspace Learning to Rank.

Function MvSL2R $(\mathbf{X}, \mathbf{Y}, k)$;
Input : The feature vectors of V views $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$, the relevance $\mathbf{y}$, and the dimensionality in the subspace k.
Output: The predicted relevance probabilities $p = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_V\}$ of the new data.
Train a neural network to update the low-dimensinonal representation representation $\mathbf{Z}_v$ e.g. in (12) and (17). and the projection matrix $\mathbf{W} = [\mathbf{W}_1 \mathbf{W}_2 \ldots \mathbf{W}_V]^\top$.
Train a logistic regressor (1) as the scoring function to obtain the weight matrix $\mathbf{a}$.
Predict the new sample pairs for their relevance probabilities using (7) with the trained sub-networks $\mathcal{F}$ and $\mathcal{G}$, and the obtained weights $\mathbf{W}$ and $\mathbf{a}$.

3.2 Multi-view Canonically Correlated Auto-Encoder (MvCCAE)

In contrast to DMvCCA and DMvMDA, where the nonlinear correlation between multiple views is optimized, we propose a multi-objective solution by maximizing the between-view correlation while minimizing the reconstruction error from each view source. Given the data matrix $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ of V views, the encoding network $\mathcal{F}$ and the decoding network $\mathcal{G}$, and the projection matrix $\mathbf{W}$, the objective of MvCCAE is formulated as follows,

$$\mathcal{J}_{\text{MvCCAE}} = \arg\max\, \mathcal{J}'_{\text{DMvCCA}} - \alpha\sum_{v}^{V}\ell_{\text{AE}}\Big(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))\Big), \tag{8}$$

where we introduce the new objective

$$\mathcal{J}'_{\text{DMvCCA}} = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\max}\frac{\text{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j=1\\ j\neq i}}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_j^\top\mathbf{W}_j\right)}{\text{Tr}\,(\sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_i^\top\mathbf{W}_i)}, \tag{9}$$

and the loss function of the vth autoencoder is $\ell_{\text{AE}}(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))) = \|\mathbf{X}_v - \mathcal{G}_v(\mathcal{F}_v(\mathbf{X}_v))\|_2 + \rho\sum_l \|\nabla_{\mathbf{X}_v}\mathcal{F}_v^l(\mathbf{X}_v)\|_2$, with the L_2 regularization at the lth intermediate layer of the vth view denoted by $\mathbf{Z}_v^l = \mathcal{F}_v^l(\mathbf{X}_v)$. Here, α and ρ are controlling parameters for the trade-off between the terms.

3.2.1 Optimization

Following the objective of DMvCCA [5], we aim to directly optimize the trace ratio in (9) and let

$$f = \text{Tr}\left(\sum_{i=1}^{V}\sum_{\substack{j\neq i\\ j=1}}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_j^\top\mathbf{W}_j\right),$$

and

$$g = \text{Tr}\left(\sum_{i=1}^{V}\mathbf{W}_i^\top\mathbf{Z}_i\,\mathbf{L}\,\mathbf{Z}_i^\top\mathbf{W}_i\right).$$

Here, the output of each sub-network $\mathcal{F}_v$ is denoted by $\mathbf{Z}_v = \mathcal{F}_v(\mathbf{X}_v)$. Then, we have

$$\frac{\partial f}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathbf{Z}_j \, \mathbf{L}, \tag{10}$$

and

$$\frac{\partial g}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V} \mathbf{W}_i \, \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}. \tag{11}$$

By using (10) and (11) and following the quotient rule, we derive the stochastic optimization of MvCCAE to be

$$\frac{\partial \mathcal{J}_{\text{MvCCAE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g \, \frac{\partial f}{\partial \mathbf{Z}_v} - f \, \frac{\partial g}{\partial \mathbf{Z}_v}\right) - \frac{\partial}{\partial \mathbf{Z}_v} \alpha \sum_{v}^{V} \ell_{\text{AE}}(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))). \tag{12}$$

The gradient to compute the autoencoding loss ℓ_{AE} is derived from the view-specific sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$. The sub-network $\mathcal{F}_v$ is optimized with $\frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v}$ to obtain the output $\mathbf{Z}_v$, while the gradient of $\mathcal{G}_v$ network with respect to its parameters can be obtained using the chain rule from $\frac{\partial \mathcal{G}_v(\mathbf{X}_v)}{\partial \mathbf{Z}_v}$.

3.3 Multi-view Modularly Discriminant Auto-Encoder (MvMDAE)

Similar to MvCCAE, the objective of MvMDAE is to optimize the combination of the view-specific reconstruction error and the cross-view correlation as follows,

$$\mathcal{J}_{\text{MvMDAE}} = \arg\max \mathcal{J}'_{\text{DMvMDA}} - \alpha \sum_{v}^{V} \ell_{\text{AE}}\Big(\mathbf{X}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))\Big). \tag{13}$$

The new objective for the cross-view correlation is

$$\mathcal{J}'_{\text{DMvMDA}} = \underset{\mathbf{W}^\top \mathbf{W} = \mathbf{I}}{\arg\max} \frac{\text{Tr}\,(\sum_{i=1}^{V}\sum_{j=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_B \mathbf{Z}_j^\top \mathbf{W}_j)}{\text{Tr}\,(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \mathbf{L}_W \mathbf{Z}_i^\top \mathbf{W}_i)}, \tag{14}$$

3.3.1 *Optimization*

The detailed optimization is derived by replacing the laplacian matrix in MvCCAE with $\mathbf{L}_B$ and $\mathbf{L}_W$ in (14). We let

$$f = \text{Tr}\left(\sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_B \, \mathbf{Z}_j^\top \mathbf{W}_j\right),$$

and

$$g = \text{Tr}\left(\sum_{i=1}^{V} \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_W \, \mathbf{Z}_i^\top \mathbf{W}_i\right).$$

Then, we have

$$\frac{\partial f}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V} \sum_{\substack{j \neq i \\ j=1}}^{V} \mathbf{W}_i \, \mathbf{W}_j^\top \mathbf{Z}_j \, \mathbf{L}_B, \tag{15}$$

and

$$\frac{\partial g}{\partial \mathbf{Z}_i} = \sum_{i=1}^{V} \mathbf{W}_i \, \mathbf{W}_i^\top \mathbf{Z}_i \, \mathbf{L}_W. \tag{16}$$

The stochastic optimization of MvMDAE can be derived by using (15), (16) and applying the quotient rule as follows,

$$\frac{\partial \mathcal{J}_{\text{MvMDAE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g \, \frac{\partial f}{\partial \mathbf{Z}_v} - f \, \frac{\partial g}{\partial \mathbf{Z}_v}\right) - \frac{\partial}{\partial \mathbf{Z}_v} \alpha \sum_{v}^{V} \ell_{\text{AE}}(\mathbf{X}_v, \mathcal{G}_v(\mathcal{F}_v(\cdot))). \tag{17}$$

The gradient of the objective can be calculated using the chain rule, and the stochastic gradient descent (SGD) is used with mini-batches for optimization.

3.4 Deep Multi-view Discriminant Ranking (DMvDR)

Multi-view Subspace Learning to Rank provides a promising method with MvCCAE and MvMDAE. However, it does not have a direct connection to ranking. Continuing the idea of multi-objective optimization, we propose to optimize the view-specific and the joint ranking together in the single network as shown in Fig. 2. Taking university ranking as an example, various ranking lists are generated from different agencies, and each agency uses a different set of attributes to represent the universities. In training, given the inputs $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$, the cross entropy loss (2) is optimized with the view-specific relevance $\mathbf{y}$ and the joint view relevance $\overline{\mathbf{y}}$. Based on their evaluation metrics, the attributes $\mathbf{X}_v$, where $v = 1, \ldots, V$, are trained through the view-specific sub-network $\mathcal{F}_v$. The nonlinear representations $\mathbf{Z}_v = \mathcal{F}_v(\mathbf{X}_v), v = 1, \ldots, V$, are the inputs of the joint network $\mathcal{H}$ as $\mathbf{W}_v^\top \mathbf{Z}_v$, $v = 1, \ldots, V$, after the mappings to generate the joint university ranking list. Each of them is also the input to the view-specific network $\mathcal{G}_v$, which minimizes its distance to the original ranking $\mathbf{r}_v$. We similarly exploit the effectiveness of intermediate layers $\mathbf{Z}_v$ in between the view-specific sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$, but towards the ranking loss for DMvDR. The detailed procedure of this method is described below.

The gradient of each view-specific sub-network $\mathcal{G}_v$ is calculated from the output $\mathbf{y}$ with respect to its parameters. Since the loss passes from each view-specific $\mathcal{F}_v$ to $\mathcal{G}_v$ sub-network, the gradient can be calculated independently with respect to the output of each view-specific $\mathcal{F}_v$ sub-network as $\frac{\partial \mathbf{y}}{\partial \mathbf{Z}} = \{\frac{\partial \mathbf{y}_1}{\partial \mathbf{Z}_1}, \frac{\partial \mathbf{y}_2}{\partial \mathbf{Z}_v}, \ldots, \frac{\partial \mathbf{y}_V}{\partial \mathbf{Z}_V}\}$ Then, the gradient of $\frac{\partial \mathbf{y}_v}{\partial \mathcal{G}_v}$ with respect to its network weights can be determined through backpropagation [46]. All sub-networks contain several layers with Sigmoid functions.

The fused sub-network $\mathcal{H}$ is updated with the gradient of the ranking loss from the cross-view relevance scores $\overline{\mathbf{y}}$. Similar to the generation of training data in MvSL2R, we find the intersection of the ranking data with different representations or measurements from various sources, and perform the pairwise transform to have the sample pairs as the input $\mathcal{X}$ and $\overline{\mathbf{y}}$ from the cross-view ranking orders $\overline{\mathbf{r}}$ in (6). As a result, the input $\mathbf{S}$ to the fused sub-network $\mathcal{H}$

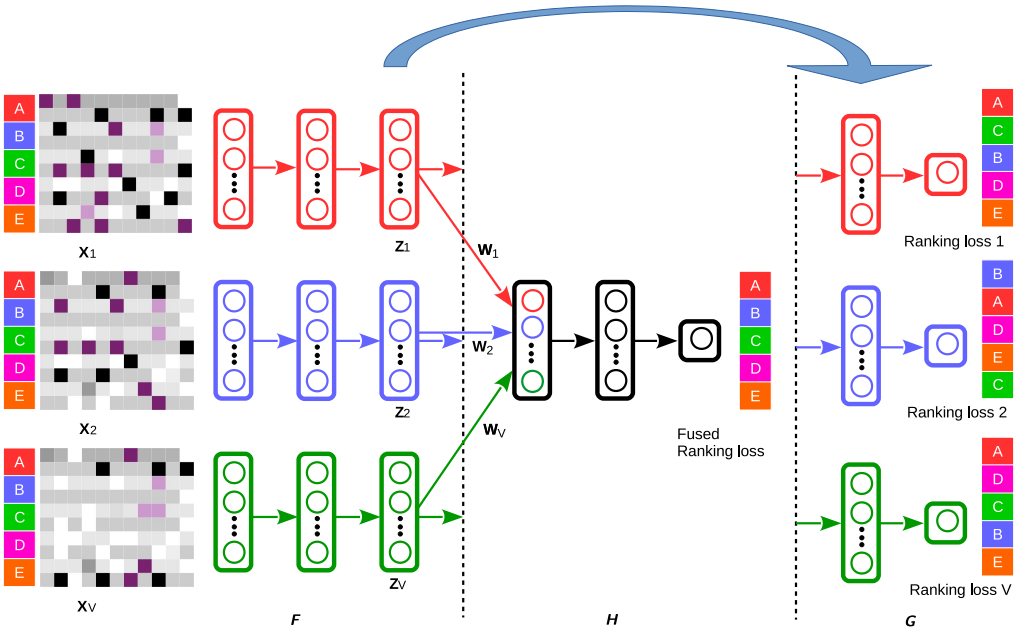


Fig. 2: System diagram of the Deep Multi-view Discriminant Ranking (DMvDR). First, the features $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V\}$ are extracted for data representations in different views and fed through the individual sub-network $\mathcal{F}_v$ to obtain the nonlinear representation $\mathbf{Z}_v$ of the vth view. The results are then passed through two pipelines of networks. One line goes to the projection $\mathbf{W}$, which maps all $\mathbf{Z}_v$ to the common subspace, and their concatenation is trained to optimize the fused ranking loss with the fused sub-network $\mathcal{H}$. The other line connects $\mathbf{Z}_v$ to the sub-network $\mathcal{G}_v, \forall v = 1, \ldots, V$ for the optimization of the vth ranking loss.

is the concatenation of the nonlinear mapping from the V view-specific networks $\mathcal{F}_v$ as

$$\mathbf{S} = [\mathbf{W}_1^\top \mathbf{Z}_1 \;\; \mathbf{W}_2^\top \mathbf{Z}_2 \;\; \ldots \;\; \mathbf{W}_V^\top \mathbf{Z}_V]^\top. \tag{18}$$

In testing, we can distinguish two possible scenarios: (a) If the samples are aligned and all presented from each view, the results from nonlinear mappings are combined in the same manner as the training phase to generate a fused ranking list $\overline{\mathbf{p}}$ at the end of $\mathcal{H}$ sub-network; and (b) If there are missing samples or completely unaligned in the test data, $\mathbf{S} = \mathbf{W}_v^\top \mathbf{Z}_v$ for the vth view. The resulting view-specific prediction $\mathbf{p}_v$ still maintains the cross-view agreement which is ranked from the trained joint network. The gradient of $\frac{\partial \overline{\mathbf{y}}}{\partial \mathbf{S}}$ and $\frac{\partial \overline{\mathbf{y}}}{\partial \mathcal{H}}$ can be easily calculated afterwards using the SGD.

Joint ranking is achieved using a multi-view subspace embedding layer. Similar to MvMDAE, we take the mappings from the outputs from the sub-networks $\mathcal{F}_v$. The gradient of multi-view subspace embedding (MvSE) in the trace ratio form is calculated by combining (15) and (16):

$$\frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} = \frac{1}{g^2}\left(g \frac{\partial f}{\partial \mathbf{Z}_v} - f \frac{\partial g}{\partial \mathbf{Z}_v}\right). \tag{19}$$

The embedding layer is important as its gradient is forward passed to the fused sub-network $\mathcal{H}$. Meanwhile, it is backward propagated in the layers of $\mathcal{F}_v$ to reach the input $\mathbf{X}_v$. In turn, the parameters in $\mathcal{G}_v$ are also affected by the outputs of $\mathcal{F}_v$ sub-networks.

The update of the view-specific $\mathcal{F}_v$ depends on the view-specific ranking output $\mathbf{y}$ and the cross-view relevance $\overline{\mathbf{y}}$ as it is a common sub-network in both pipelines of networks. Through backpropagation, the v-th sub-networks $\mathcal{F}_v$ and $\mathcal{G}_v$ are optimized consecutively with respect to the gradient $\frac{\partial \mathbf{y}}{\partial \mathbf{X}_v}$. Meanwhile, the training error with respect to the fused ranking $\overline{\mathbf{y}}$ is passed through multi-view subspace embedding (MvSE) from $\mathbf{S}$ in (18) as the input to the fused sub-network $\mathcal{H}$. The resulting gradient of each sub-network $\mathcal{F}_v$ is given by

$$\begin{aligned}\frac{\partial \mathcal{J}_{\text{DMvDR}}}{\partial \mathbf{Z}_v} = &\frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} - \alpha \sum_v^V \frac{\partial}{\partial \mathbf{Z}_v} \ell_{\text{Rank}}(\mathbf{X}_v, \mathbf{y}_v; \mathcal{G}_v(\mathcal{F}_v(\cdot))) \\ &- \beta \frac{\partial}{\partial \mathbf{Z}_v} \ell_{\text{Rank}}(\mathbf{S}, \overline{\mathbf{y}}; \mathcal{H}(\cdot)),\end{aligned} \tag{20}$$

where α and β are the scaling factors controlling the magnitude of the ranking loss. Similar to the other sub-networks, the gradients with respect to their parameters can be obtained by following the chain rule.

The update of the entire network of DMvDR can be summarized using the SGD with mini-batches. The parameters of the sub-network are denoted by $\theta = \{\theta_{\mathcal{F}_1}, \theta_{\mathcal{F}_2}, \ldots, \theta_{\mathcal{F}_V}, \theta_{\mathcal{G}_1}, \theta_{\mathcal{G}_2}, \ldots, \theta_{\mathcal{G}_V}, \theta_{\mathcal{H}}\}$. A gradient descent step is $\Delta\theta = -\eta \frac{\partial}{\partial \theta} \mathcal{J}_{\text{DMvDR}}$, where η is the learning rate. The gradient update step at time t can be written down

with the chain rule collectively:

$$\begin{aligned}
\Delta\theta^t &= \{\Delta\theta^t_{\mathcal{F}_1}, \Delta\theta^t_{\mathcal{F}_2}, \ldots, \Delta\theta^t_{\mathcal{F}_V}, \\
&\quad \Delta\theta^t_{\mathcal{G}_1}, \Delta\theta^t_{\mathcal{G}_2}, \ldots, \Delta\theta^t_{\mathcal{G}_V}, \Delta\theta^t_{\mathcal{H}}\} \\
\Delta\theta^t_{\mathcal{G}_v} &= -\frac{\partial \ell_{\text{rank}}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathcal{G}_v} \\
\Delta\theta^t_{\mathcal{H}} &= -\frac{\partial \ell_{\text{rank}}}{\partial \overline{\mathbf{y}}} \cdot \frac{\partial \overline{\mathbf{y}}}{\partial \mathcal{H}} \\
\Delta\theta^t_{\mathcal{F}_v} &= \frac{\partial \mathcal{J}_{\text{MvSE}}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v} - \frac{\partial \ell_{\text{rank}}}{\partial \mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v} \\
&\quad - \frac{\partial \ell_{\text{rank}}}{\partial \overline{\mathbf{y}}} \cdot \frac{\partial \overline{\mathbf{y}}}{\partial \mathbf{S}} \cdot \frac{\partial \mathbf{S}}{\partial \mathbf{Z}_v} \cdot \frac{\partial \mathbf{Z}_v}{\partial \mathcal{F}_v}.
\end{aligned} \tag{21}$$

We generate the training data using the pairwise transform presented in Section 2.1.1. The weights are normalized to the unit norm during backpropagation. In testing, the test samples can also be transformed into pairs to evaluate the relative relevance of each sample to its query. The raw ranking data can also be fed into the trained model to predict their overall ranking positions.

4 Experiments

In this section, we evaluate the performance of the proposed multi-view learning to rank methods in three challenging problems: university ranking, multi-linguistic ranking and image data ranking. The proposed methods are also compared to the related subspace learning and co-training methods. The subspace learning methods follow the steps proposed in Section 3.1 for ranking. All neural network topologies are chosen based on a validation set and trained for 100 epoches. We compare the performance of the following methods in the experiments:

- **Best Single View**: a method which shows the best performance of Ranking SVM [27] over the individual views.
- **Feature Concat**: a method which concatenate the features of the common samples for training a Ranking SVM [27].
- **LMvCCA [5]**: a linear multi-view CCA method.
- **LMvMDA [5]**: a linear supervised method for multi-view subspace learning.
- **MvDA [6]**: another linear supervised method for multi-view subspace learning. It differs from the above in that the view difference is not encoded in this method.
- **SmVR [8]**: a semi-supervised method that seeks a global agreement in ranking. It belongs to the category of co-training. We develop the complete data in the following experiments for training so that its comparison with the subspace learning methods is fair. Therefore, SmVR becomes a supervised method in this paper.
- **DMvCCA [5]**: a nonlinear extension of LMvCCA using neural networks.
- **DMvMDA [5]**: a nonlinear extension of LMvMDA using neural networks.
- **MvCCAE**: the first proposed multi-view subspace learning to rank method proposed in the paper.
- **MvMDAE**: the supervised multi-view subspace learning to rank method proposed in the paper.
- **DMvDR**: the end-to-end multi-view learning to rank method proposed in the paper.

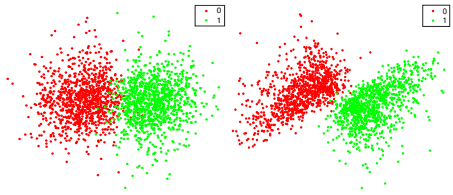
We present the quantitative results using several evaluation metrics including the Mean Average Precision (MAP), classification accuracy and Kendal's tau. The Average Precision (AP) measures the relevance of all query and sample pairs with respect to the same query, while the MAP score calculates the mean AP across all queries [47]. After performing pairwise transform on the ranking data, the relevance prediction can be considered as a binary classification problem, and therefore the classification is utilized for evaluation. Kendal's tau measures the ordinal association between two lists of samples.

We also present the experimental results graphically, and the following measures are used. The Mean Average Precision (MAP) score, which is the average precision at the ranks where recall changes, is illustrated on the 11-point interpolated precision-recall curves (PR curve) to show the ranking performance. Also, the ROC curve provides a graphical representation of the binary classification performance. It shows the true positive rates against the false positive rate at different thresholds. The correlation plots show linear correlation coefficients between two ranking lists.

4.1 University Ranking

The university ranking dataset available in Kaggle.com [48] collects the world ranking data from three rating agencies, including the Times Higher Education (THE) World University Ranking, the Academic Ranking of World Universities (ARWU), and the Center for World University Rankings (CWUR). Despite political and controversial influences, they are widely considered as authorities for university ranking. The measurements are used as the feature vectors after feature preprocessings, which includes feature standardization and removal of categorical variables and groundtruth indicators including the ranking orders, university name, location, year and total scores. The 271 common universities from 2012 to 2014 are considered for training. After the pairwise transform in each year, 36542 samples are generated as the training data. The entire data in 2015 is considered for testing. The data distribution (after binary transform) of the 196 common universities in 2015 is shown in Fig. 3. We use a topology of [16 32] in the hidden layers of $\mathcal{F}_v$ for both MvCCAE and MvMDAE, and the final layer has 10 dimensions. The decoding network of each view has 64 hidden sigmoid nerons before reconstructing to the input. In DMvDR, $\mathcal{F}_v$ consists of [50 10] hidden neurons, and both the vth decoding network $\mathcal{G}_v$ and $\mathcal{H}$ have 100 neurons in the hidden layer.

We can make several observations from the data distribution in Fig. 3. Firstly, the pairwise transform is applied on the university ranking data, which equally assigns the original ranking data to two classes. Then, the dimensionality of the data is reduced to 2-dimensional using PCA in order to display it on the plots of Fig. 3. The data is then labelled with two colors red and green indicating the relevance between samples. We can notice a high overlap



(a) Raw data distribution. (b) Projected data distribution.

Fig. 3: The left plot shows the data distribution by concatenating the measurements as features of the common universities from 3 different agencies in 2015. The right plot shows the concatenated and projected features using MvMDAE for the same universities.

between the two classes in the case of raw data (left plot of Fig. 3), while the data on the right is clearly better separated after the projection using the proposed MvMDAE. This shows the discrimination power of the proposed supervised embedding method.

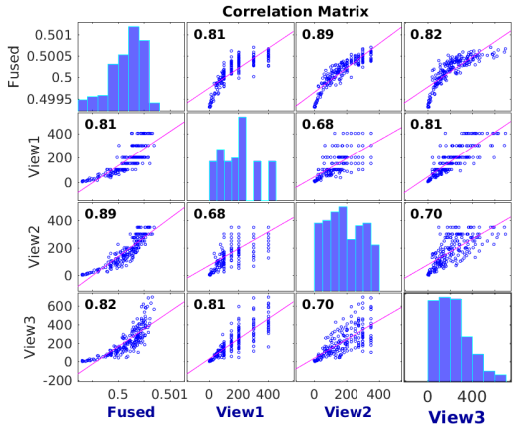


Fig. 4: Rank correlation matrix for views 1-3 and the fused view.

Furthermore, a rank correlation matrix of plots is presented in Fig. 4 with correlations among pairs of ranking lists from the views 1-3 and the predicted list denoted by 'Fused'. Histograms of the ranking data are shown along the matrix diagonal, while scatter plots of data pairs appear off diagonal. The slopes of the least-squares reference lines in the scatter plots are equal to the displayed correlation coefficients. The fused ranking list is produced by the proposed DMvDR, and the results are also generated from the common universities in 2015. We first take a closer look at the correlations between the views 1-3. The correlation coefficients are generally low, with the highest (0.81) between view 1 and 3, while the others are around 0.70. In contrast, the fused rank has a high correlation to each view. The scatter plots and the reference lines are well aligned, and the correlation coefficients are all above 0.80, demonstrating that the proposed DMvDR effectively exploits the global agreement with all view.

Finally, the average prediction results over 3 different university datasets of the proposed and competing methods are reported in Table 1. Due to the misalignment of ranking data in 2015 across datasets, we make the ranking prediction based on each view input, which is further elaborated in the Section 3.4. We observe that Ranking SVM [27] on the single feature or its concatenation performs poorly compared to the other methods. This shows that when the data is heterogeneous, simply combining the features cannot enhance joint ranking. Kendal's tau from the linear subspace learning methods are comparatively higher than their nonlinear counterparts. This is due to the fact that the nonlinear methods aim to maximize to the correlation in the embedding space, while the scoring function is not optimized for ranking. In contrast, DMvDR optimizes the entire ranking process, which is confirmed with the highest ranking and classification performance.

TABLE 1: Average Prediction Results (%) on 3 University Ranking Datasets in 2015.

Methods	Kendal's tau	Accuracy
Best Single View	65.38	-
Feature Concat	35.10	-
LMvCCA [5]	86.04	94.49
LMvMDA [5]	87.00	94.97
MvDA [6]	85.81	94.34
SmVR [8]	80.75	-
DMvCCA [5]	70.07	93.20
DMvMDA [5]	70.81	94.75
MvCCAE (*ours*)	75.94	94.01
MvMDAE (*ours*)	81.04	94.85
DMvDR (*ours*)	**89.28**	**95.30**

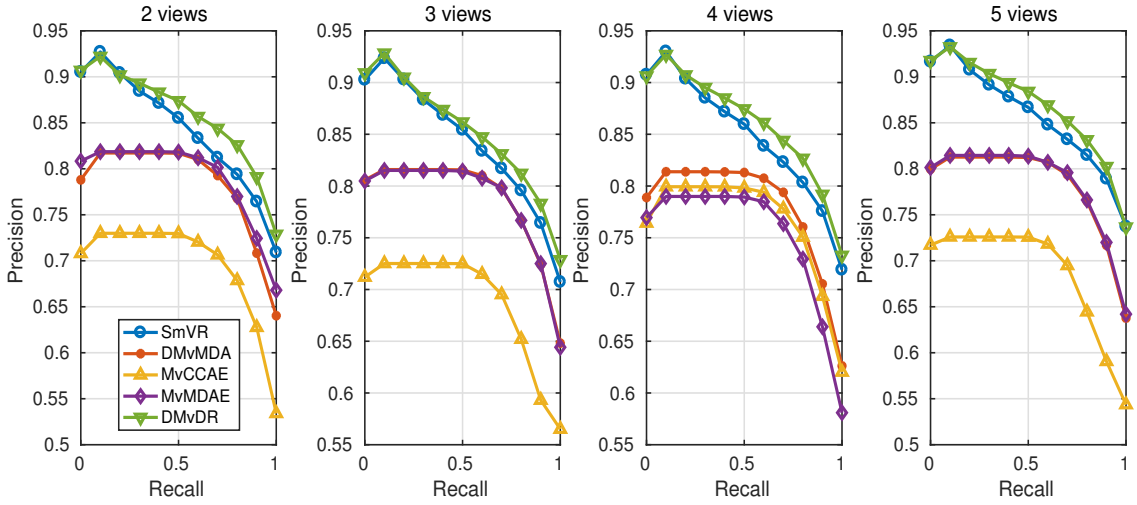
4.2 Multi-lingual Ranking

The Multi-lingual Ranking is performed on Reuters RCV1/RCV2 Multi-lingual, Multi-view Text Categorization Test collection [3]. We use Reuters to indicate this dataset in later paragraphs. It is a large collection of documents with news ariticles written in five languages, and grouped into 6 categories by topic. The bag of words (BOW) based on a TF-IDF weighting method [47] is used to represent the documents. The vocabulary has a size of approximately 15000 on average and is very sparse.

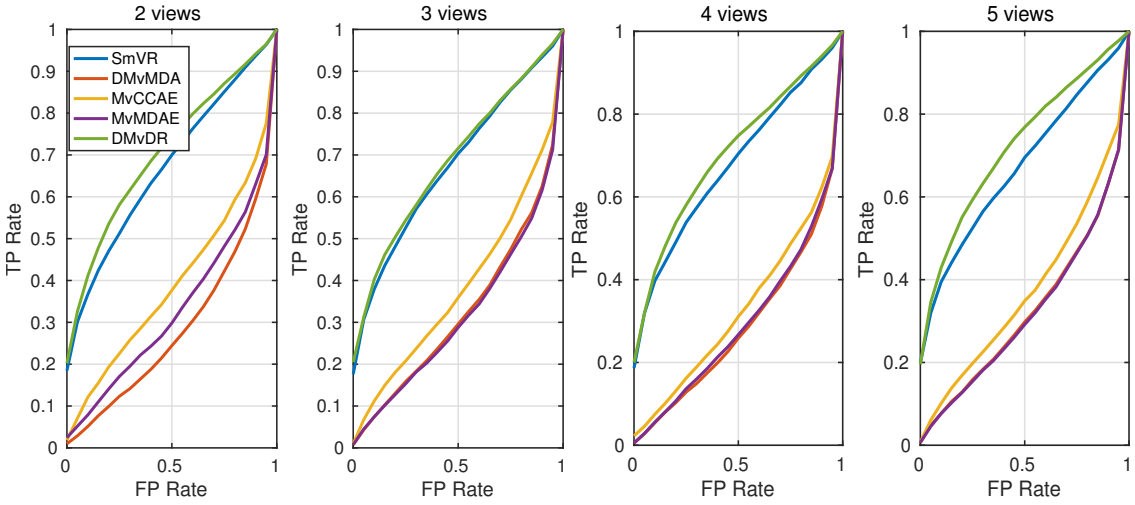
We consider the English documents and their translations to the other 4 languages in our experiment. Specifically, the 5 views are numbered as follows:

- View 1: original English documents;
- View 2: English documents translated to French;
- View 3: English documents translated to German;
- View 4: English documents translated to Italian;
- View 5: English documents translated to Spanish.

Due to its high dimensionality, the BOW representation of each document is projected using a sparse SVD to a 50-dimensional compact feature vector. We randomly select 40 samples from each category in each view as training data. The training data composed of 28680 samples is generated



(a) PR curve on Reuters.



(b) ROC curve on Reuters.

Fig. 5: The PR and ROC curves with 2-5 views applied to Reuters dataset.

TABLE 2: Quantitative Results (%) on the Reuter Dataset.

	2 views		3 views		4 views		5 views	
Methods	MAP@100	Accuracy	MAP@100	Accuracy	MAP@100	Accuracy	MAP@ 100	Accuracy
Feature Concat	58.87	70.41	56.97	70.10	57.59	69.88	58.46	69.97
LMvCCA [5]	59.10	70.20	62.40	72.01	54.41	66.61	60.41	**72.62**
LMvMDA [5]	59.09	70.16	58.81	71.94	61.54	72.45	59.28	72.07
MvDA [6]	55.95	69.03	55.42	67.57	55.64	68.64	58.93	68.46
SmVR [8]	78.37	71.44	78.24	71.15	78.66	71.37	79.36	71.64
DMvCCA [5]	53.87	67.41	42.68	62.02	54.51	68.03	57.27	65.00
DMvMDA [5]	60.08	71.40	63.12	70.93	61.55	72.12	62.52	70.78
MvCCAE (*ours*)	48.75	66.43	49.10	62.90	60.70	71.86	48.80	63.05
MvMDAE (*ours*)	62.63	**74.20**	63.02	71.04	60.74	72.60	62.74	71.20
DMvDR (*ours*)	**80.01**	72.68	**79.34**	**72.23**	**80.32**	**73.07**	**81.64**	72.39

between pairs of English documents based on the pairwise transform in Section 2.1.1, and the translations to other languages are used for augmenting the views. We select another 360 samples from 6 categories and create a test dataset of 64620 document pairs. If considering the ranking function linear as proved in [24], we make document pairs comparable and balance them by assigning some of the data to the other class with the opposite sign of the feature vectors, so that the number of samples is equally distributed in both classes. For MvCCAE and MvMDAE, the encoding network $\mathcal{F}_v$ has the topology of [100 10 10], while the decoding network $\mathcal{G}_v$ has a hidden layer of 32 neurons. $\mathcal{F}_v$ in DMvDR has the topology of [50 10], while $\mathcal{G}_v$ has sigmoid neurons in the structore of [64 1]. $\mathcal{H}$ is a sub-network of [100 1].

We first analyze the PR and ROC curves in Fig. 5. Since we have all translations of the English documents, each sample is well aligned in all views and, therefore we perform joint learning and prediction in all multi-lingual experiments. The experiments start with 2 views with English and its translation to French, and then the views are augmented with the documents of other languages. Subspace ranking methods are trained by embedding with increasing number of views, while SmVR as a co-training takes two views at a time, and the average performance of all pairs is reported. The proposed methods with two competing ones are included in the plots in Fig. 5. The proposed DMvDR clearly performs the best across all views as can be seen in the PR and ROC plots in Fig. 5. SmVR is the second best with a lower precision and less area under curve compared to DMvDR. Among the remaining three methods, DMvMDA performs favorably in the PR curves but not as well in the ROC plots. The results are comparatively consistent across all views.

We can observe the quantitative MAP and accuracy results in Table 2. It shows that the linear methods together with the feature concatenation have similar results which are generally inferior to the nonlinear methods in classification. Note also that nonlinear subspace learning methods cannot provide any superior MAP scores, which can be explained by the fact that the embedding is only intended to construct a discrimative feature space for classifying the pairs of data. We can also observe the MAP scores and accuracies are stable across views. This can be interpreted as the global ranking agreement can be reached to a certain level when all languages correspond to each other. It is again confirmed that the end-to-end solution consistently provides the highest scores, while SvMR is a few percentages behind. When the features from different views follow a similar data distribution, the co-training method performs well and competes with the proposed DMvDR.

4.3 Image Data Ranking

Image data ranking is a problem to evaluate the relevance between two images represented by different types of features. We adopt the Animal With Attributes (AWA) dataset [49] for this problem due to its diversity of animal appearance and large number of classes. The dataset is composed of 50 animal classes with a total of 30475 images, and 85 animal attributes. We follow the feature generation in [5] to adopt 3 feature types forming the views:

- Image Feature by VGG-16 pre-trained model: a 1000-dimensional feature vector is produced from each image by resizing them to 224×224 and taken from the outputs of the $fc8$ layer with a 16-layer VGGNet [50].
- Class Label Encoding: a 100-dimensional Word2Vector is extracted from each class label. Then, we can map the visual feature of each image to the text feature space by using a ridge regressor with a similar setting as in [5] to genenate another set of textual feature, with connection to the visual world. The text embedding space is constructed by training a skip-gram [51] model on the entire English Wikipedia articles, including 2.9 billion words.
- Attibute Encoding: an 85-dimensional feature vetor can be produced with a similar idea as above. Since each class of animals contains some typical patterns of the attribute, a 50×85 lookup table can be constructed to connect the classes and attributes [52], [53]. Then, we map each image feature to the attribute space to produce the mid-level feature.

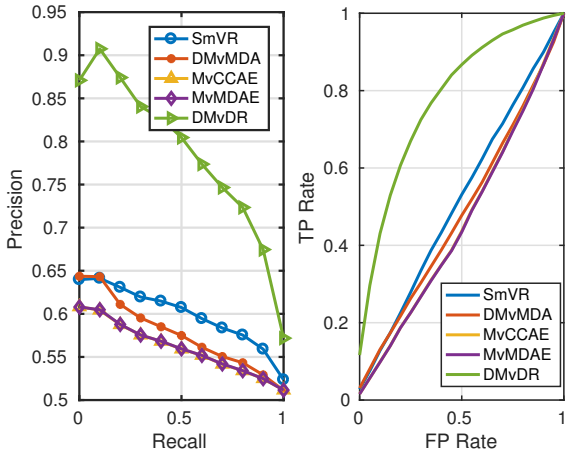


Fig. 6: PR and ROC curves on AWA.

We generate the image ranking data as follows. From the 50 classes of animal images, we find 400 pairs of images with 200 in-class pairs and 200 out-of-class image pairs from each class. We then end up with 20000 training data pairs. Similarly, we will have 20000 test data pairs. We select 40 images from each class used for training data and a separate set of 40 samples as test data. Another 10 images are used as queries: 5 of them are associated with the in-class images as positive sample pairs and 5 as negative sample pairs. For the negative sample pairs, we randomly select 40 classes from 49 remaining animal classes at a time, and one image per class is associated with each query image under study. We found a common topology for both MvCCAE and MvMDAE from the validation set. The encoding networks have the topology of [64 10] in the hidden layers with sigmoid neurons and the output dimensionality in the

common latent space is 10. The decoding networks have one hidden layer with 50 sigmoid neurons and each has a final layer of the input dimensionality. For DMvDR, each of the encoding networks $\mathcal{F}_v$ has a topology of [100 100 10] and its vth decoding netuwork $\mathcal{G}$ has [100 1] sigmoid neurons. The merging network $\mathcal{H}$ has also the network structure of [100 1]. $\alpha = 0.01$ is chosen from the grid search.

TABLE 3: Quantitative Results (%) on the AWA Dataset.

Methods	MAP@100	Accuracy
Feature Concat	38.08	50.60
LMvCCA [5]	49.97	51.85
LMvMDA [5]	49.70	52.35
MvDA [6]	49.20	52.82
SmVR [8]	52.12	50.33
DMvCCA [5]	51.38	50.83
DMvMDA [5]	51.52	51.38
MvCCAE (*ours*)	49.01	53.28
MvMDAE (*ours*)	48.99	53.30
DMvDR (*ours*)	**76.83**	**71.48**

We can observe the performance of the methods on the animal dataset graphically in Fig. 6 and quantitatively in Table 3. DMvDR outperforms the other competing methods by a large margin as shown in the plots of Fig. 6. Due to the variety of data distribution from different feature types as view inputs, the co-training type of SmVR can no longer compete with the end-to-end solution. From Table 3, one can observe that the performance of the feature concatenation suffers from the same problem. On the other hand, our proposed subspace ranking methods produces satisfactory classification rates while the precisions remain somewhat low. This implies again the scoring function is critical to be trained together with the feature mappings. The other linear and nonlinear subspace ranking methods have comparatively similar performance at a lower position.

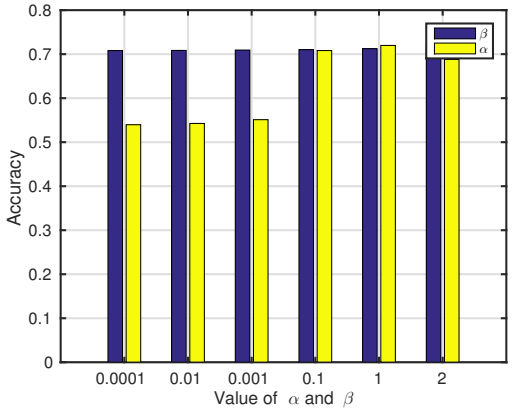


Fig. 7: Performance of DMvDR on different values of α and β.

We also study the effects of selecting different values of α and β during training on the classification performance of DMvDR. The results are shown in Fig. 7. One parameter varies across the neural network models while the other one is fixed based on a grid search. While the performance is mostly consistent with the value changes of α and β, it drops when α is below 0.001. It shows the importance of jointly optimizing the view-specific ranking.

5 Conclusion

Learning to rank has been a popular research topic with numerous applications, while multi-view ranking remains a relatively new research topic. In this paper, we aimed to associate the multi-view subspace learning methods with the ranking problem and proposed three methods in this direction. MvCCAE is an unsupervised multi-view embedding method, while MvMDAE is its supervised counterpart. Both of them incorporate multiple objectives, with a correlation maximization on one hand, and reconstruction error minimization on the other hand, and have been extended in the multi-view subspace learning to rank scheme. Finally, DMvDR is proposed to exploit the global agreement while minimizing the individual ranking losses in a single optimization process. The experimental results validate the superior performance of DMvDR compared to the other subspace and co-training methods on multi-view datasets with both homogeneous and heterogeneous data representations.

In the future, we will explore the scenario when there exists missing data, which is beyond the scope of the current proposed subspace ranking methods during training. Multiple networks can also be combined by concatenating their outputs, and further optimized in a single sub-network. This solution may also be applicable for homogeneous representations.

References

[1] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[2] Y. Zhu, G. Wang, J. Yang, D. Wang, J. Yan, J. Hu, and Z. Chen, "Optimizing search engine revenue in sponsored search," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 588–595.

[3] M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views - an application to multilingual text categorization," in *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., 2009, pp. 28–36.

[4] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[5] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, 2017, doi: 10.1109/TCYB.2017.2742705.

[6] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 38, no. 1, pp. 188–194, Jan 2016.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[8] N. Usunier, M.-R. Amini, and C. Goutte, "Multiview semi-supervised learning for ranking multilingual documents," *Machine Learning and Knowledge Discovery in Databases*, pp. 443–458, 2011.

[9] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[10] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 2010, pp. 1–4.

[11] A. Iosifidis, A. Tefas, and I. Pitas, "Kernel reference discriminant analysis," *Pattern Recognition Letters*, vol. 49, pp. 85–91, 2014.

[12] A. Iosifidis and M. Gabbouj, "Nyström-based approximate kernel subspace learning," *Pattern Recognition*, vol. 57, pp. 190–197, 2016.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1247–1255.

[14] G. Cao, A. Iosifidis, and M. Gabbouj, "Multi-view nonparametric discriminant analysis for image retrieval and recognition," *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1537–1541, Oct 2017.

[15] F. Feng, L. Nie, X. Wang, R. Hong, and T.-S. Chua, "Computational social indicators: A case study of chinese university ranking," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2017, pp. 455–464.

[16] "The times higher education world university ranking," https://www.timeshighereducation.com/world-university-rankings, 2016.

[17] N. C. Liu and Y. Cheng, "The academic ranking of world universities," *Higher education in Europe*, vol. 30, no. 2, pp. 127–136, 2005.

[18] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Picture collage," *IEEE Transactions on Multimedia (TMM)*, vol. 11, no. 7, pp. 1225 –1239, 2009.

[19] X. Li, T. Pi, Z. Zhang, X. Zhao, M. Wang, X. Li, and P. S. Yu, "Learning bregman distance functions for structural learning to rank," *IEEE Transactions on Knowledge and Data EngineeringS*, vol. 29, no. 9, pp. 1916–1927, Sept 2017.

[20] O. Wu, Q. You, X. Mao, F. Xia, F. Yuan, and W. Hu, "Listwise learning to rank by exploring structure of objects," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1934–1939, 2016.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[22] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems*, 1998, pp. 451–457.

[23] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 89–96.

[24] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," 2000.

[25] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *The Journal of machine learning research*, vol. 4, pp. 933–969, 2003.

[26] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[27] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[28] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[29] J. Xu and H. Li, "AdaRank: a boosting algorithm for information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 391–398.

[30] C. J. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions," in *Advances in neural information processing systems*, 2007, pp. 193–200.

[31] H.-J. Ye, D.-C. Zhan, Y. Miao, Y. Jiang, and Z.-H. Zhou, "Rank consistency based multi-view learning: a privacy-preserving approach," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 991–1000.

[32] W. Gao and P. Yang, "Democracy is good for ranking: Towards multi-view rank learning and adaptation in web search," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 63–72.

[33] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[34] M. Borga, "Canonical correlation: a tutorial," http://people.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf, 2001.

[35] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing (TIP)*, vol. 11, no. 3, pp. 293–305, 2002.

[36] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[37] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, Nov 2015.

[38] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[39] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[40] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 1083–1092.

[41] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational neural networks," *Neural computation*, 2016.

[42] P. Nousi and A. Tefas, "Deep learning algorithms for discriminant autoencoding," *Neurocomputing*, vol. 266, pp. 325–335, 2017.

[43] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[44] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 1, pp. 40–51, 2007.

[45] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[46] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 1998, pp. 9–50.

[47] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[48] "World university rankings: A kaggle dataset." https://www.kaggle.com/mylesoneill/world-university-rankings, 2016.

[49] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 36, no. 3, pp. 453–465, 2014.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[51] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NIPS)*, 2013, pp. 3111–3119.

[52] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, vol. 3, 2006, p. 5.

[53] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith, "Default probability," *Cognitive Science*, vol. 15, no. 2, pp. 251–269, 1991.

Guanqun Cao received the double B.Eng. degree in Electronic and Information/Computer Engineering from Huazhong University of Science and Technology, China and University of Birmingham, UK. He also received the M.Sc degree from the joint Erasmus Mundus programme in Color Informatics and Media Technology. He is currently a PhD student at the Multimedia Research Group, Tampere University of Technology, Finland. His research interests include multimedia retrieval and machine learning with a focus on multi-view data analysis.

Alexandros Iosifidis (SM'16) received the Diploma and M.Eng. degrees in Electrical and Computer Engineering from the Democritus University of Thrace, Greece, in 2008 and 2010, respectively, and the Ph.D. degree in Informatics from the Aristotle University of Thessaloniki, Greece, in 2014. He was a Postdoctoral Researcher in Aristotle University of Thessaloniki and Tampere University of Technology from 2015 until 2017. He was a recipient of a two-year Post-Doctoral Fellowship Award given by TUT Foundation to five (in total) Post-Doctoral Researchers in all scientific fields in 2015 and the Academy of Finland Postdoctoral Research Fellowship in 2016. He was the recipient of the H.C. Orsted Forskerspirer prize in 2018 for research excellence at a young age. He is currently an Assistant Professor of Machine Learning and Computer Vision in the Department of Engineering, Aarhus University, Denmark.

Dr. Iosifidis is a Senior Member of IEEE and he served as an Officer of the Finnish IEEE Signal Processing/Circuits and Systems Chapter from 2016 to 2018. He also serves as an Associate Editor in Neurocomputing and IEEE Access journals and Area Editor in Signal Processing: Image Communications journal. His current research interests lie in the areas of Machine Learning and Pattern Recognition, with applications mainly in images/videos and time series data. He has co-authored 48 journal papers, 70 conference papers, and 4 book chapters in topics of his expertise.

Moncef Gabbouj (F'11) received his BS degree in electrical engineering in 1985 from Oklahoma State University, and his MS and PhD degrees in electrical engineering from Purdue University, in 1986 and 1989, respectively. Dr. Gabbouj is a Professor of Signal Processing at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Academy of Finland Professor during 2011-2015. His research interests include multimedia content-based analysis, indexing and retrieval, machine learning, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Dr. Gabbouj is a Fellow of the IEEE and member of the Academia Europaea and the Finnish Academy of Science and Letters. He is the past Chairman of the IEEE CAS TC on DSP and committee member of the IEEE Fourier Award for Signal Processing. He served as associate editor and guest editor of many IEEE, and international journals and Distinguished Lecturer for the IEEE CASS. He organized several tutorials and special sessions for major IEEE conferences and EUSIPCO. Dr. Gabbouj guided 46 PhD students and published 700 papers.

Vijay Raghavan is the Alfred and Helen Lamson Endowed Professor in Computer Science at the Center for Advanced Computer Studies. His research interests are in information retrieval and extraction, data and web mining, multimedia retrieval, data integration, and literature-based discovery. He has published around 275 peer-reviewed research papers These and other research contributions cumulatively accord him an h-index of 37, based on Google Scholar citations to his publications. He has served as major adviser for 29 doctoral students and has garnered over $\$13$ million in external funding. Dr. Raghavan brings substantial technical expertise, interdisciplinary collaboration experience, and management skills to his projects.

Dr. Raghavan's service work at the university includes coordinating the Louis Stokes-Alliance for Minority Participation (LS-AMP) program since 2001. Raghavan has served as PC Chair, PC Co-chair or PC member in countless ACM and IEEE conferences. He received the ICDM 2005 Outstanding Service Award. Raghavan was an ACM National Lecturer from 1993 to 2006. He was a member of the Advisory Committee of the NSF Computer and Information Science and Engineering directorate, from 2008-2010. He serves on the Executive Committee of the IEEE -TCII and on the Steering committees of WIC Consortium and Int'l Rough Sets society. He received WIC 2013 Outstanding Service Award. He is one of the Editors-in-Chief of Web Intelligence journal and an Associate Editor of ACM Transactions on Internet Technology.

Dr. Raghavan is the founding director of the Visual and Decision Informatics (CVDI), an NSF-funded Industry University Cooperative Research Center, which started its phase II (second 5 years) operations in March 2017, and is a co-director of the Laboratory for Internet Computing.

Raju Gottumukkala is Director of Research for the Informatics Research Institute at UL Lafayette. He leads several research initiatives in the broader area of cyber physical systems and big data across three centers within IRI. This includes The National Science Foundation (NSF) Center for Visual and Decision Informatics (CVDI) and Center for Business and Information Technologies. He is also the Site Director of NSF CVDI - an NSF Industry University Cooperative Research Center in the area of big data.

Dr. Gottumukkala's research interest is in the broader area of cyber physical systems - specifically addressing real-world informatics and integrated systems modelling issues. He has led various efforts in the area of big data platforms, system resilience, modelling $\&$ verification of distributed systems, software defined networks, visual analytics, and evolutionary networks. His research has generated over $\$7M$ in funding from various state and federal agencies including NSF, DHS S$\&$T, DOE, state agencies, and the private sector. He has 15 peer-reviewed conference/journal publications, 2 U.S. Patents, and authored several technical reports. He is also part of the US Ignite community leadership group and representative from the city of Lafayette. He has also served on various conference programs, and review committees.